УДК 621.314.58 DOI: 10.24027/2306-7039.1.2017.102000

ИНФОРМАЦИОННЫЙ АНАЛИЗ РЕГРЕССИОННОЙ МОДЕЛИ ФАКТОРНОГО ПЛАНИРОВАНИЯ МЕДИКО-БИОЛОГИЧЕСКОГО ЭКСПЕРИМЕНТА

П.Ф. Щапов, *доктор технических наук, профессор Национального технического университета "Харьковский политехнический институт" (НТУ "ХПИ")*
Р.С. Томашевский, *кандидат технических наук, доцент НТУ "ХПИ", г. Харьков*
Е.В. Зольтман, *аспирант НТУ "ХПИ", г. Харьков*

П.Ф. Щапов

Р.С. Томашевский

Е.В. Зольтман

Приведено описание измерительного научного медико-биологического эксперимента для решения задачи идентификации состояния биологического объекта по биопотенциальным случайным измерительным сигналам.

In paper there is given the description of the measurement scientific medical and biological experiment for solving the problem of identification of the condition of biological object according to the biopotential occasional measurement signals.

Одним из типичных диагностических тестов в медицине является идентификация состояния биологического объекта (образца) по результатам измерений значений откликов (изменений параметров) этого объекта при воздействии на него нормированного физического или химического фактора [1–3]. К таким исследованиям могут быть отнесены: биохемилюминесцентный анализ плазмы крови, активируемая колориметрия мочи, миография и электрокардиография под нагрузкой, электроэнцефалографические и кожногальванические исследования реакции на стимулирующие факторы (стресс-тесты). Для корректного решения такой задачи необходимо определение степени воздействия такого фактора на биологический объект (образец), функциональной связи между уровнями фактора (факторной нагрузки) и значениями параметров объекта. Планирование такого (активного) научного эксперимента — это сложная многовариантная и многокритериальная задача, результативность которой зависит от выбора условий, обязательных к выполнению и обеспечивающих максимальную эффективность эксперимента, количественно оцениваемую в рамках заданных статистических требований [4]. Любой измерительный эксперимент с биологическими объектами проводится в условиях неустранимой неопределенности свойств таких объектов, поскольку большинство этих свойств определяется динамикой случайных процессов внутренних биохимических или биофизических изменений. Контроль и управление такими процессами практически невозможны, что переводит планирование подобных исследований на более сложный и высокий уровень, учитывающий факторную неоднородность.

Количественная теория информации достаточно хорошо развита в теории синтеза информационных систем [5]. Особенно интересным является ее использование для создания информационной теории измерений [6]. Задачи оценки количества информации при постановке диагностического теста и эффективное планирование такого эксперимента решаются во многих актуальных отраслях, таких как газотранспортная, авиационная, автомобильная, экономика предприятий и т. д. [7–9]. Существуют работы по применению подобных методов для повышения количества информации при лабораторных исследованиях в биологии и медицине. К недостатку последней теории [6] следует отнести обязательность использования ограничений на неопределенность результатов измерений при безусловной необходимости нормативного метрологического обеспечения измерительного эксперимента. Очень часто в таких исследованиях в качестве фактора выступает время, особенно при недостатке информации о функции факторной нагрузки.

Любой измерительный научный эксперимент, предназначенный для выявления количественно выраженных связей между уровнями влияющего фактора и выходной контролируемой переменной, решает задачу формализации стохастической модели такой связи. Если уровни фактора определяются достаточно точно, а неопределенность контролируемой переменной постоянна, как в случае медицинских и биофизических экспериментов, то наиболее удобной моделью стохастической связи является полиномиальная регрессия [4].

Цель работы — расширение возможностей оптимального, по максимуму ожидаемой информации, планирования научного эксперимента в задачах многопараметрической идентификации динамических объектов.

Большинство подобных экспериментов предусматривают активное прогрессирующее регулируемое изменение факторной нагрузки. Чаще всего такая нагрузка является непрерывной функцией времени эксперимента. В этом случае динамика нагрузки будет определяться последовательностью дискретных значений моментов времени измерения контролируемой величины. В работе рассматривается полиномиальная модель такой зависимости. Выходная контролируемая величина — случайный измерительный сигнал $U(t)$, изменяющийся во времени, математическое ожидание $m_U(t)$ которого является неизвестной функцией факторной нагрузки и вида состояния объекта (образца).

Функциональные особенности математического ожидания $m_U(t)$ несут информацию в виде S_0, S_1,..., S_m состояния объекта испытаний, если обеспечена стабильность и точность задаваемых во времени уровней факторной нагрузки.

Будем рассматривать две базовых задачи планирования эксперимента:

• задачу обучения информационно-измерительной системы контроля, диагностики и идентификации, связанную с восстановлением (синтезом) условных математических моделей процессов $m_U(t/S_j)$, $j=\overline{0,m}$, при экспериментальном исследовании биологических объектов с известными (верифицированными) состояниями;

• задачу контроля, диагностики или идентификации реализаций для $m_U(t)$, полученных в ходе измерительных экспериментов на объектах с априори неизвестными состояниями.

В первом случае требуется создать метрологическое и статистически представительное обеспечение для элементов сложной процедуры преобразования сигнала $U(t)$ в решения γ_0, γ_1, ..., γ_m, соответствующие идентифицируемым состояниям S_0, S_1,..., S_m объекта испытаний.

Такое обеспечение предполагает формирование функциональных стандартных моделей $\hat{m}_U(t/S_j, j=0,m)$, описывающих динамику изменений математического ожидания $m_U(t)$ на ограниченных интервалах наблюдения при выполнении обязательного условия: параметры (коэффициенты) модели должны быть статистически различными для состояний S_0, S_1,..., S_m.

Статистическая значимость такого различия должна обеспечиваться заданной достоверностью P (или рисками первого (α) и второго (β) рода принятия соответствующих неправильных решений).

Во втором случае необходимо определить условия обеспечения представительности дискретизированных и группированных значений сигнала $U(t)$ для объектов с априори неизвестными состояниями. Естественно полагать, что условия проведения обучающих (по видам состояний) и контрольного экспериментов должны быть физически идентичными.

Вероятностные модели информационной теории измерений [4, 10, 11] позволяют получить уравнение для оценки количества I ожидаемой измерительной информации для любых вариантов законов распределения случайных результатов измерения:

$$I=H_1-H_2, \tag{1}$$

где H_1 и H_2 — исходная (до измерения) и остаточная (после измерения) энтропии случайной величины, неслучайный параметр которого содержит ожидаемую информацию.

Поскольку факторный активный эксперимент в задаче идентификации состояний объекта является двухэтапным (обучение и контроль), то комплексным критерием его эффективности может служить критерий максимума ожидаемой информации результатов идентификации. Вероятностная мера такой информации I — разность исходной h_Y и остаточной $h_{Y/}y_N$ дифференциальных энтропий [12], каждая из которых, будучи увеличенной на логарифм погрешности различения Δx (на $\log_2\Delta x$, бит), образует, соответственно, либо исходную (H_1), либо остаточную (H_2) энтропию, заменяя уравнение (1) выражением

$$I=h_Y-h_{Y/yN}, \tag{2}$$

где N — количество модельных $\hat{m}_U(t)$ функций, используемых для оценивания неизвестного матожидания $m_U(t/S_j)$; Y — неизвестная, зависимая от времени t функция $m_U(t/S_j)$, определяющая нестационарность (по матожиданию) процесса $U(t)$; y_N — регрессионная усредненная по N модель $\hat{m}_U(t)$, полученная по n дискретизированным отсчетам $U(t_i)$ процесса $U(t)$, $n>>N$, $i=\overline{1,n}$.

Найдем выражение для энтропий h_Y и $h_{Y/yN}$. Рассмотрим регрессионную модель результата измерения y_i значения функции Y:

$$y_i = y_N(t_i) + \varepsilon_i; i=\overline{1,n}, \tag{3}$$

где ε_i — случайный остаток модели, вероятностные свойства которого определены условиями:

а) $M[\varepsilon_i]=0$, остаток представляет собой случайный стационарный шум;

б) $M[\varepsilon_i \cdot \varepsilon_i]=0$, для всех $i \neq e$, все отсчеты шумовой составляющей независимы;

в) $M[\varepsilon_i^2]=\sigma_\varepsilon=\text{const}$, мощность шумовой составляющей постоянна;

г) $\varepsilon_i \approx \text{NORM}(0,\sigma_\varepsilon^2)$, случайный остаток имеет нормальное распределение.

Главным является условие в однородности дисперсий остатков ε_i [4].

Усредненная модель $y_N(t_i)$ получена преобразованием

$$y_N(t_i) = N^{-1}\sum_{r=1}^{N} \hat{m}_{ur}(t_i), \quad (4)$$

где r — номер обучающего эксперимента по воспроизведению процесса $U(t)$ для фиксированного состояния S_j.

План обучаещего эксперимента должен включать, как минимум, параметры m (количество идентифицированных состояний), N (количество обучающих образцов, групп данных), n (количество измерений в каждой группе), влияние которых на величину ожидаемой информации I имеет смысл исследовать. В таком эксперименте общее количество измерений $G=(m+1)N \cdot n$.

Если модель $\hat{m}_{ur}(t_i)$ представить в виде полиномиальной регрессии степени p [5] с коэффициентами $\beta_0^*,...,\beta_p^*$

$$\hat{m}_{ur}(t_i) = \sum_{k=0}^{p} \beta_{kr}(t_i - \bar{t})^k, \quad (5)$$

то любой из β_k^* — это случайная оценка неизвестного коэффициента β_k, причем

$$\begin{cases} M[\beta_{kr}^*] = \beta_k; \\ M[(\beta_{kr}^* - \beta_k)^2] = \sigma_{\beta_k}^2. \end{cases}$$

Дисперсию $\sigma_{\beta_k}^2$ оценок β_k^* можно уменьшить за счет операции усреднения (4).

Тогда модель (3) примет вид

$$y_i = N^{-1}\sum_{r=1}^{N}\sum_{k=0}^{p} \beta_{kr}(t_i - \bar{t})^k + N^{-1}\sum_{r=1}^{N}\varepsilon_{ir}, \quad (6)$$

где свойства остатка ε_{ir} идентичны свойствам остатка ε_i модели (3), а значение y_i определено для y_N при фиксированном времени отсчета t_i.

Для нахождения дифференциальной исходной энтропии h_Y предположим, что все значения y_i лежат в интервале $[y_{max}, y_{min}]$, в границах которого лежат все y_i, соответствующие состояниям S_j, $j=\overline{1,m}$, при условии, что $m \to \infty$.

Если плотность вероятности y_i — равномерный закон по всем m состояниям, то есть

$$f(y_i / S_j) = \frac{1}{y_{max} - y_{min}},$$

если $y_i \in \Delta y$, то

$$h_Y = \int_{y_{min}}^{y_{max}} (y_{max} - y_{min})^{-1} \log(y_{max} - y_{min}) = \log(y_{max} - y_{min}). \quad (7)$$

Таким образом, энтропия h_Y отражает неопределенность появления решения γ_j, когда j может принять любое из значений интервала $[0, M]$, где $M>>m$, а сам эксперимент (обучающий или контрольный) еще не проведен.

После проведения эксперимента остаточная дифференциальная энтропия [12] имеет вид

$$h_{Y/y_N} = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y) f(y_N / y) \log\frac{f(y) f(y_N / y)}{f(y_N)} dy dy_N, \quad (8)$$

где

$$f(y) = f(y_i / S_j) = (y_{max} - y_{min})^{-1}; \quad (9)$$

$$f(y_N / y) = f(y_N - y) = f(\Delta y);$$

$f(\Delta y)$ — композиция законов распределения для величин y и $\Delta y=(y_N-y)$.

При фиксированном S_j состоянии объекта идентификации (j=const) закон распределения случайных величин y_i и Δy можно считать нормальным с дисперсиями, соответственно, σ_y^* и $\sigma_{\Delta y}^*$. Поскольку центрирование (вычитание математических ожиданий) не меняет энтропию h_{y/y_N}, нормальные законы распределения $f(y_N/y)$ и $f(y_N)$ будут иметь вид

$$f(y_N / y) = \frac{1}{\sigma_{\Delta y}\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{\Delta y}{\sigma_{\Delta y}})^2}; \quad (10)$$

$$f(y_N) = \frac{1}{\sqrt{(\sigma_y^2 + \sigma_{\Delta y}^2)2\pi}} \cdot e^{-\frac{1}{2}(\frac{y_N^2}{\sigma_y^2 + \sigma_{\Delta y}^2})}. \quad (11)$$

Выражение (8) можно рассматривать как первый начальный момент [6] величины (математическое ожидание)

$$\log\frac{f(y) f(y_N / y)}{f(y_N)}.$$

Используя знак математического ожидания, запишем

$$h_{y/y_N} = -M\left[\log\frac{f(y) f(y_N / y)}{f(y_N)}\right]. \quad (12)$$

Подставим в правую часть выражения (12) вероятностные модели (9), (10) и (11) для законов распределения $f(y)$, $f(y_N/y)$ и $f(y_N)$, получим с учетом свойств математического ожидания [6]

$$h_{y/y_N} = -\log\left\{\frac{\sqrt{\dfrac{\sigma_y^2+\sigma_{\Delta y}^2}{\sigma_{\Delta y}^2}}}{(y_{max}-y_{min})}\right\} + \log\left\{e^{\frac{M[\Delta y^2]}{2\sigma_{\Delta y}^2}+\frac{M[y_N^2]}{2(\sigma_y^2+2\sigma_{\Delta y}^2)}}\right\}, \quad (13)$$

где $\Delta y = y_N - y$.

Учитывая, что $M[\Delta y^2] = \sigma_{\Delta y}^2$, а $M[y_N^2] = \sigma_y^2 + \sigma_{\Delta y}^2$, получим для второго логарифма в правой части выражение

$$\log\left\{e^{\frac{M[\Delta y^2]}{2\sigma_{\Delta y}^2}+\frac{M[y_N^2]}{2(\sigma_y^2+2\sigma_{\Delta y}^2)}}\right\} = \log 1 = 0.$$

Дифференциальная остаточная энтропия h_{y/y_N} имеет окончательный вид

$$h_{y/y_N} = -\log\left\{\frac{\sqrt{1+\dfrac{\sigma_y^2}{\sigma_{\Delta y}^2}}}{(y_{max}-y_{min})}\right\}. \quad (14)$$

С учетом выражений (7) и (14) для дифференциальных энтропий h_y и h_{y/y_N}, количество ожидаемой информации (2) о результате идентификации запишется как

$$I = \frac{1}{2}\log(y_{max}-y_{min}) - \left[-\log\left\{\frac{\sqrt{1+\dfrac{\sigma_y^2}{\sigma_{\Delta y}^2}}}{(y_{max}-y_{min})}\right\}\right].$$

С учетом логарифмических преобразований, получим

$$I = \frac{1}{2}\log\left(1+\frac{\sigma_y^2}{\sigma_{\Delta y}^2}\right). \quad (15)$$

Дисперсия σ_Y^2 — это дисперсия равномерно распределенной величины y с плотностью распределения вероятностей (9) [6]:

$$\sigma_y^2 = \frac{(y_{max}-y_{min})^2}{12}. \quad (16)$$

Для нахождения дисперсии $\sigma^2_{\Delta y}$ воспользуемся линейным относительно случайных величин β^*_{kr} и ε_{ir} преобразованием (6), задавая для этих случайных величин их дисперсии $\delta^2_{\beta k}$ и δ^2_{ε} соответственно [6]:

$$\sigma^2_{\Delta y_i} = N^{-2}\sum_{r=1}^{N}\sum_{k=0}^{p}\delta^2_{kr}(t_i-\bar{t})^{2k} + \frac{\delta^2_{\varepsilon}}{N}. \quad (17)$$

Подставляя выражения для дисперсий σ_y^2, $\sigma^2_{\Delta y_i}$ в уравнения (16), (17), получим

$$I_i = \frac{1}{2}\log\left\{1+\frac{(y_{max}-y_{min})^2}{12[N^{-2}\sum_{r=1}^{N}\sum_{k=0}^{p}\delta^2_{\beta_{kr}}(t_{ir}-\bar{t}_r)^{2k}+\dfrac{\delta^2_{\varepsilon}}{N}}\right\}. \quad (18)$$

Для анализа полученного выражения (15) и определения условий планирования эксперимента по идентификации неизвестного состояния S_j объекта испытаний упростим полученное выражение, используя две последовательные процедуры усреднения для произведения $\delta^2_{\beta_{kr}}(t_i-\bar{t})^{2k}$:

$$\begin{cases}\delta^2_{\beta r}(t_{ir}-\bar{t}_r)^2 = \dfrac{1}{(p+1)}\sum\limits_{k=0}^{p}\delta_{\beta_{kr}}(t_{ir}-\bar{t}_r)^{2k};\\ \delta^2_{\beta}(t_i-\bar{t})^2 = \dfrac{1}{N}\sum\limits_{r=1}^{N}\delta^2_{\beta r}(t_i-\bar{t}_r)^2.\end{cases} \quad (19)$$

С учетом (19) выражение (18) примет окончательный (для последующего анализа) вид:

$$I_i = \frac{1}{2}\log\left\{1+\frac{(y_{max}-y_{min})^2}{12[\delta^2_{\beta}(t_i-\bar{t})^2+\delta^2_{\varepsilon}(p+1)^{-1}]}\left(\frac{N}{p+1}\right)\right\}. \quad (20)$$

Поскольку полная информация I удовлетворяет требованию аддитивности [12],

$$I = \sum_{j=1}^{m}\sum_{i=1}^{n} I_i,$$

то, с учетом измерений, выражение (20) примет следующий вид:

$$I = \frac{1}{2}\sum_{j=0}^{m}\sum_{i=1}^{n}\log\left\{1+\frac{(y_{max}-y_{min})^2}{12\left[\delta^2_{\beta}\left(t_i-\bar{t}\right)^2+\delta^2_{\varepsilon}\left(p+1\right)^{-1}\right]}\cdot\left(\frac{N}{p+1}\right)\right\}. \quad (21)$$

Из выражения (21) следует, что план контрольного эксперимента, помимо параметров m, N и n, будет содержать параметр p, зависящий в рамках модели (21) от параметра N, что указывает на необходимость объединения всех параметров в плане (m, N, n, p).

Условия, выполнение которых приводит к увеличению количества информации, получаемой по i-му дискретизированному отсчету процесса $U(t)$, а следовательно, к снижению рисков идентификации вида состояния объекта (или, что то же самое, повышению достоверности идентификации), можно сформулировать следующим образом.

1. Широкая шкала для идентифицируемых состояний, то есть надо увеличивать разность $\left(y_{max}-y_{min}\right)$ при увеличении количества m таких состояний.

2. Увеличение числа p информативных параметров $(\beta_0, \beta_1, ..., \beta_p)$ должно проводиться с соблюдением условия неуменьшения отношения

$$\theta = \frac{N}{p+1}.$$

Это означает, что увеличение p должно сопровождаться ростом N объема стандартизированных выборок для процесса $U(t)$.

3. Желательно стремиться к увеличению количества стандартных обучающих образцов N при фиксированных значениях p числа информативных параметров, обеспечивая условие $\theta \geqslant \text{const}$.

4. Дисперсии $\delta^2_{\beta r}$ полиномиальных коэффициентов можно уменьшить, приближая момент отсчета t_i к математическому ожиданию времени наблюдения $\bar{t}$.

Последнее указывает, что:

1) отсчеты процесса $U(t_i)$ вблизи $\bar{t}$ более информативны, чем отсчеты процесса на периферии (относительно $\bar{t}$) его наблюдения;

2) имеет смысл нормирование процесса $U(t)$ по времени (в пределах периода его наблюдения при однократном воздействии факторной нагрузки).

5. Желательно увеличивать степень p регрессионного полинома, учитывая при этом условие (4) (в этом случае уменьшается дисперсия случайного остатка δ^2_{ε}).

6. В любом случае, желательно увеличивать значение θ, повышая объем N при фиксированных значениях p, m плана обучающего и контрольного экспериментов.

7. Желательно увеличивать число n дискретизированных отсчетов, исследуемых для формирования усредненной регрессионной модели (4). Это не только повышает значение полной информации (с учетом ее аддитивности), но и уменьшает дисперсии $\delta^2_{\beta k}$, $k = \overline{0, p}$ оценок угловых коэффициентов β^*_k регрессионной модели (5), что базируется на свойстве состоятельности этих оценок [5].

8. Из выражения (21) следует, что полная информация растет с увеличением m числа диагностируемых состояний.

Список литературы

1. Amplifier design for EMG recording from stimulation electrodes during functional electrical stimulation leg cycling ergometry / R. Shalaby, T. Schauer, W. Liedecke, J. Raisch // Biomed Tech. – New York, 2011; 56: 23–33.
2. Новый способ диагностики состояния клеток человека с помощью электрохимических биосенсоров / В. Н. Ослопов, Ю. В. Ослопова, Д. В. Сайфуллина [и др.] // Вестник современной клинической медицины. – 2012. — Т. 5, вып. 3. – С. 12–15.
3. Effects of 10 Hz and 20 Hz transcranial alternating current stimulation (tACS) on motor functions and motor cortical excitability / C. Wach, V. Krause, V. Moliadze [et al.] // Behav Brain Res. – 2013. – 241, 1–6.
4. *Джонсон Н.* Статистика и планирование эксперимента в технике и науке: Методы планирования эксперимента / Н. Джонсон, Ф. Лион; пер. с англ. под ред. Э. К. Лецкого. — М.: Мир, 1981. — 520 с.
5. *Себер Дж.* Линейный регрессионный анализ / Дж. Себер; пер. с англ. под ред. М. Б. Малютова. — М.: Мир, 1980. — 456 с.
6. *Вентцель Е. С.* Теория вероятностей и её инженерные приложения: учеб. пособие для вузов / Е. С. Вентцель, Л. А. Овчаров. — М.: Высшая школа, 2000. — 480 с.
7. *Гишваров А. С.* Оценка состояния газотурбинного привода с использованием информационного потенциала статистических данных эксплуатации / А. С. Гишваров, Е. А. Могильницкий, И. И. Гиззатуллин // Авиационная и ракетно-космическая техника. – Уфа: УГАТУ, 2012. – Т. 16, № 5 (50). – С. 13–19.
8. *Васин Н. С.* Комплексное диагностирование деятельности предприятия / Н. С. Васин // Вопросы экономики. – 2013. – 48 (186).
9. *Машошин О. Ф.* Диагностика авиационной техники: учеб. пособие / О. Ф. Машошин. — М.: МГТУ ГА, 2007. — 141 с.
10. *Королюк В. С.* Справочник по теории вероятностей и математической статистике / В. С. Королюк, Н. И. Портенко; под ред. В. С. Королюка. — К.: Наукова думка, 1978. — 584 с.
11. *Кисіль І. С.* Метрологія, точність і надійність засобів вимірювань: навч. посібник / І. С. Кисіль. — Івано-Франківськь: Факел, 2000. — 400 с.
12. *Орнатский П. П.* Теоретические основы информационно-измерительной техники / П. П. Орнатский. — К.: Вища школа, 1983. — 455 с.