КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2011 Т. 3 № 1 С. 93–99

Ки&М

АНАЛИЗ И МОДЕЛИРОВАНИЕ СЛОЖНЫХ ЖИВЫХ СИСТЕМ

УДК: 57.087

Оптимизация интегральных оценок состояния биосистем с использованием параллельных вычислений

Ю. В. Кистенев[1], О. Ю. Никифорова[2], Г. Г. Стромов[1], В. А. Фокин[1,a]

[1] ГОУ ВПО «Сибирский государственный медицинский университет» Минздравсоцразвития России, медико-биологический факультет
Россия, 634050, г. Томск, ул. Московский тракт, д. 2
[2] Институт оптики атмосферы им. В. Е. Зуева СО РАН
Россия, 634021, г. Томск, пл. Академика Зуева, д. 1

E-mail: [a] fokin@ssmu.ru

Получено 28 февраля 2011 г.

В работе рассмотрен подход к оптимизации интегральных оценок состояния биосистем. Подход включает процедуры уменьшения вариабельности интегральных оценок, основанные на статистическом моделировании экспериментальных данных, а также оптимизацию числа признаков состояния на основе оценки их относительного вклада в интегральную оценку с использованием параллельных вычислений.

Ключевые слова: интегральная оценка состояния биосистем, информативность показателей, статистическое моделирование

Optimization of integral estimation of bio-systems state using parallel calculation

Yu. V. Kistenev[1], O. Yu. Nikiforova[2], G. G. Stromov[1], V. A. Fokin[1]

[1] Siberian State Medical University, 2, Moscowski Trakt st., Tomsk, 634050, Russia
[2] V. E. Zuev Institute of Atmospheric Optics SB RAS, 1, Academician Zuev square, Tomsk, 634021, Russia

Abstract. – The approach to optimization of integral estimation of bio-systems state is presented. The approach is included the procedures of decreasing of variability of integral estimation based on statistical modeling of experimental data set and optimization the quantity of a state characteristics on a base of their relative contribution to the integral estimation using parallel calculation.

Keywords: integral estimation of bio-systems, self-descriptiveness characteristics, statistical modeling

Citation: *Computer Research and Modeling*, 2011, vol. 3, no. 1, pp. 93–99 (Russian).

Работа выполнена при частичной финансовой поддержке РФФИ (грант № 09-02-99038 р_офи), ФЦПК (госконтракт № 02.740.11.0083).

Интегральная оценка близости объекта к эталону состояния

Разработка методов оценки состояния биосистем на основе анализа медико-биологических данных представляет собой активно развивающееся направление [Диагностика состояния человека, 2003; Генкин, 1999; Дюк, Эммануэль, 2003; Armitage, Berry, 1994]. Решаемая проблема связана с тем, что при изменении состояния биосистемы вариации показателей признаков состояния зачастую носят разнонаправленный характер, в ряде случаев оставаясь практически в пределах статистических норм. Основные подходы к решению данной проблемы заключаются в следующем.

Выбор в качестве признаков состояния небольшого набора независимых показателей. При этом оценкой состояния системы служит величина самого непосредственно измеряемого показателя. Однако для сложных систем обнаружить показатели, которые бы однозначно определяли ее состояние, бывает достаточно трудно.

Формирование оценок состояния с использованием методов многомерного статистического анализа данных. Сюда можно отнести всевозможные процедуры многомерного регрессионного, дискриминантного, факторного и других многомерных методов анализа. Достаточно эффективно для проведения интегральных оценок последовательное применение нескольких статистических процедур анализа [Миронкина, Бобров, 1998; Подвальный, Матасов, Бырко, 2002]. При этом в качестве оценок состояния используются характеристики, следующие из соответствующего метода статистического анализа. Основной проблемой использования методов многомерного статистического анализа данных является то, что они предъявляют достаточно жесткие требования к объему и методам формирования выборок, которые не всегда выполнимы в условиях медико-биологического исследования.

Использование обобщенных показателей, формируемых на основе анализа закономерностей функционирования изучаемых биосистем. Это подходы, связанные с построением полуэмпирических индексов состояния, таких как биохимический, клинический индексы; получением оценок функционального состояния организма (уровень функционирования, функциональный резерв, степень напряженности регуляторных механизмов); анализом вербальных данных на основе теории нечетких множеств, применением методов многомерного шкалирования, нейросетевых технологий и т. п. [Нисевич, Марчук, 1982; Баевский, 1989; Муха, Скворцов, Авдеюк, 2003]. Их можно рассматривать как модельно-независимые оценки, характеризующие систему в целом.

Задача интегральной оценки состояния биосистемы сводится к выбору признаков $(y_1, y_2, ..., y_n)$ состояния и построению функционального отображения их значений в одномерную шкалу. Удобно производить оценку состояния биосистемы S по отношению к ее эталону состояния S_0. В качестве последнего для человека может быть выбрано, например, состояние здорового организма.

Пусть состояния биосистемы S_0 и S представлены наборами объектов $\{\vec{x}_i \,|\, i \in N_{S_0}\}$ и $\{\vec{y}_j \,|\, j \in N_S\}$, где N_{S_0} и N_S – объемы выборок. При этом величина интегральной оценки близости объекта $\vec{y}_j \in S$ к эталону состояния может быть определена следующим образом [Фокин, 2007]:

$$I_{S_0}(\vec{y}_j) = \frac{d(\vec{y}_j, S_0)}{D_{S_0}}, \tag{1}$$

где $d(\vec{y}_j, S_0)$ – мера близости объекта $\vec{y}_j$ к эталонному состоянию S_0; D_{S_0} – мера компактности области, занимаемой в пространстве признаков объектами, относящимися к S_0.

Мера близости объекта $\vec{y}_j$ к эталонному состоянию S_0 рассчитывалась по формуле:

$$d(\vec{y}_j, S_0) = \frac{1}{N_{S_0}} \sum_{i=1}^{N_{S_0}} d(\vec{y}_j, \vec{x}_i), \tag{2}$$

где $d(y_j, x_i)$ – расстояние Махаланобиса между объектами, вычисление которого опирается на использование матрицы ковариации признаков, что позволяет учесть взаимозависимость признаков, характеризующих состояние биосистемы [Конрадов, 1994].

Параметр D_{S_0} зададим в виде внутримножественного расстояния [Ту, Гонсалес, 1978, с. 266]:

$$D_{S_0} = \frac{1}{N_{S_0}} \sum_{i=1}^{N_{S_0}} \frac{1}{N_{S_0} - 1} \sum_{l=1}^{N_{S_0}-1} d(\vec{x}_i, \vec{x}_l), \quad (3)$$

т. е., как усредненное значение средних расстояний от каждого объекта, относящегося к состоянию S_0, до всех оставшихся. Нормировка на величину D_{S_0} в выражении (1) позволяет учесть как конфигурацию области S_0, так и взаимное расположение объектов в ней.

Уменьшение вариабельности интегральной оценки

Основная проблема практического применения формул (1)–(3) к интегральной оценке состояния биосистемы обусловлена малыми объемами выборок и широкой внутри- и междуиндивидуальной вариабельностью значений признаков объектов, характеризующих эталонное состояние, что приводит к значительному разбросу интегральных оценок.

Решение данной проблемы может быть реализовано методами статистического моделирования, которые позволяют сформировать модельные референтные выборки, обеспечивающие получение устойчивых оценок состояния. Моделирование целесообразно проводить в два этапа. Сначала моделируется M выборочных множеств X_k ($k = \overline{1, M}$) заданного объема, соответствующих статистическим характеристикам референтного состояния S_0, представленного некоторым выборочным множеством объектов $X : \{\vec{x}_i \mid i \in \overline{1, N_{S_0}}\}$. Полученные последовательности значений имитируют взятие выборок из одной и той же совокупности и, следовательно, будут свободны от погрешностей, обусловленных влиянием внутри- и междуиндивидуальной вариабельности данных. Далее, для каждого множества X_k вычисляются величины оценок $I_{S_0,k}(\vec{y})$. Здесь вектор $\vec{y}$ характеризует объект, для которого производится оценка.

Проиллюстрируем данный подход на примере интегральной оценки состояния системы красной крови по данным сканирующей электронной микроскопии (СЭМ) поверхностной архитектоники эритроцитов крови. Известно, что при онкозаболеваниях различной природы обнаруживаются изменения метаболического статуса циркулирующих эритроцитов, происходят структурно-функциональные перестройки в их мембранах, наблюдается появление трансформированных форм клеток и других нарушений [Козинец, Быкова, Сукиасова, 1982; Эритроциты и злокачественные образования, 2000].

В качестве вектора состояния нами были использованы 12 показателей, характеризующие процентное содержание различных видов переходных, предгемолитических и дегенеративных форм эритроцитов, полученных по результатам обследований больных при различных локализациях онкологических заболеваний II–III стадий, а также здоровых лиц (данные получены коллективом авторов под руководством академика РАМН, проф. В. В. Новицкого). Эталонное состояние было представлено показателями СЭМ красной крови, измеренными у 46 здоровых лиц. Оцениваемые состояния были представлены следующим количеством пациентов: рак легких – 64, рак желудка – 27, рак тонкой кишки – 23, рак области головы и шеи – 27.

Анализ экспериментальных данных показал, что при различных локализациях онкологического заболевания практически для всех измеренных показателей СЭМ наблюдаются статистически значимые их отклонения от показателей эталонного состояния, однако дать количест-

венную оценку степени изменений происходящих в системе красной крови по исходным данным не представлялось возможным.

Статистическое оценивание I_{S_0} проводилось путем моделирования выборок, соответствующих объемам N_{S_0}, равным 50, 100, 200, 400, 600, 800 и 1000 наблюдений, в предположении, что данные эталонной выборки удовлетворяют многомерному нормальному закону распределения. Каждая выборка моделировалась от 100 до 1000 раз с шагом 100. Результаты моделирования интегральной оценки состояния системы красной крови по данным СЭМ при различных локализациях рака для некоторых значений N_{S_0} и $M = 500$ представлены на рис. 1.

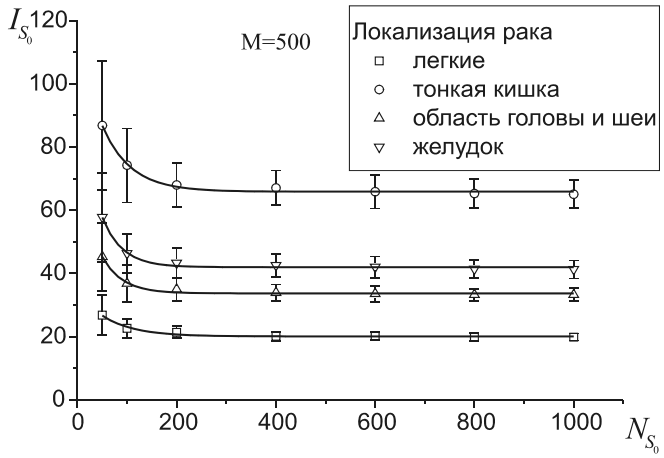

Рис. 1. Зависимость величины интегрального критерия I_{S_0} от объема модельной выборки N_{S_0} референтного состояния

Представленные результаты показывают, что на получение устойчивых оценок существенно влияет величина объема эталонной выборки N_{S_0}: коэффициент вариации при $N_{S_0} = 50$ составляет в среднем 20…25% для всех рассматриваемых состояний, уменьшаясь до 4…8% при объемах выборок $N_{S_0} = 1000$. Для других объемов модельных выборок зависимости имеют аналогичный вид. Отрезками указаны 95% доверительных интервалов.

Применение предлагаемого подхода позволило проранжировать степень изменений в системе красной крови в зависимости от локализации опухолевого процесса. В частности, при раке органов пищеварительной системы (желудок, тонкая кишка) рассчитанная величина интегральной оценки степени изменений в системе красной крови оказалась значительно выше, чем при раке легкого или опухолях головы и шеи, что соответствует имеющимся представлениям о влиянии этих патологий на систему красной крови [Эритроциты и злокачественные образования, 2000].

Оптимизация интегральной оценки

Другая проблема оптимизации интегральных оценок связана с выбором набора показателей, характеризующих состояние объекта. Так, возможны следующие варианты выбора показателей для получения интегральной оценки состояния.

1. Использование всех измеренных показателей, описывающих определенный уровень структурно-функционального описания системы.

2. Использование набора показателей, формируемых на основе экспертных заключений специалистов предметной области.

3. Использование набора показателей, наиболее информативных по какому-либо критерию оптимизации.

При использовании первого варианта, наряду с информативными, в анализ будет включаться и ряд показателей, не несущих полезной информации, что может привести к неоптимальным оценкам состояния. Второй вариант более эффективен, однако он может быть реализован лишь для ограниченного числа состояний, обеспеченных экспертной информацией.

С учетом вышесказанного нами был выбран третий вариант. В качестве критерия оптимизации использовался относительный вклад показателя в интегральную оценку состояния.

Для выделения наиболее информативных показателей проводился расчет интегральных оценок $I_{S_0}^{x_i}(\vec{x})$ на множестве измеренных показателей $\vec{x}$ после исключения из него показателя x_i. Критерием информативности S_{x_i} исключенного показателя считалось относительное изменение значения интегральной оценки биосистемы:

$$S_{x_i} = \frac{\left| I_{S_0}(\vec{x}) - I_{S_0}^{x_i}(\vec{x}) \right|}{I_{S_0}(\vec{x})} \cdot 100\%. \qquad (4)$$

Ранжирование набора показателей по информативности проводилось по следующему критерию:

$$S_{\{x_i\}} \xrightarrow[\{x_i\}]{} \max. \qquad (5)$$

Поиск информативных показателей в такой методике требует больших затрат машинного времени, поскольку число различных вариантов расчета равно числу сочетаний C_n^k, где n – размерность вектора состояния, k – количество информативных показателей. Например, при $n = 12,\ k = 3$ количество вариантов расчета равно 440.

С учетом этого для решения оптимизационной задачи поиска наиболее информативных показателей оценки состояния был разработан алгоритм параллельных вычислений интегральной оценки, реализованный на вычислительном кластере СКИФ Cyberia, расположенном в Томском государственном университете.

В данном алгоритме распараллеливания была использована методика параллелизма задач. Она подразумевает, что вычислительная задача разбивается на несколько относительно самостоятельных подзадач, выполняемых на отдельном процессоре. Такой выбор обусловлен тем, что критерий оптимизации (5) основан на проведении циклических вычислений, различающихся только изменением наборов входных данных, и не требует организации взаимодействия и синхронизации между отдельными вычислительными процессами. Алгоритм параллельных вычислений состоит из трех этапов.

1. Подготовка файлов входных данных эталонного и оцениваемого состояний для каждого из вариантов расчетов, соответствующих заданным подмножествам показателей.

2. Проведение расчетов для каждого из вариантов параллельно на отдельных узлах кластера.

3. Сбор результатов расчета в отдельный выходной файл.

Ниже представлен пример использования данной методики оптимизации интегральной оценки для анализа активизации газовыделений колонии бактерий M. Smegmatis, выращенной на мясопептонном агаре (МПА), при воздействии на колонию раствором NaCl различных концентраций. Компонентами вектора признаков $\vec{x}$, описывающего состояние объекта, являлись значения 12 коэффициентов поглощения газовыделений колоний бактерий в диапазоне частот 931–952 $см^{-1}$ генерации перестраиваемого CO_2-лазера. Проводилось измерение по 5 сканов спектра через 2 и 4 суток без добавления и с добавлением 1%, 5% и 10% раствора NaCl (оцениваемые состояния). Эталонное состояние было представлено вектором коэффициентов поглощения газовыделений чистого МПА (60 сканов спектра).

Расчеты проводились при следующих параметрах: объем моделируемой эталонной выборки – 1 000 сканов; количество моделируемых выборок – 600. При выбранных параметрах коэффициент вариации полученных оценок составил 2÷4%.

В таблице 1 представлены результаты оценки информативности коэффициента поглощения на каждой из частот диапазона 931–952 см$^{-1}$ в различные сроки и при различных концентрациях NaCl культивации бактерий M. Smegmatis, рассчитанные по критерию (5).

Таблица 1. Оценка информативности S_{λ_i} (%) коэффициентов поглощения газовыделений колоний бактерий на отдельных частотах диапазона 931–952 см$^{-1}$

λ_i см$^{-1}$	NaCl , %							
	Нет		1%		5%		10%	
	2 сут.	4 сут	2 сут.	4 сут	2 сут.	4 сут	2 сут.	4 сут
931,01	25,2	27,7	25,8	26,0	33,8	23,8	27,6	28,6
932,92	59,7	62,9	63,1	65,6	48,3	56,0	40,0	50,4
934,93	56,6	59,4	58,3	63,6	45,7	53,8	34,1	44,7
936,77	59,9	62,9	63,1	65,7	44,3	60,5	42,2	49,9
938,7	65,6	68,5	67,9	70,6	50,0	64,9	44,9	54,6
940,56	53,6	55,5	52,9	59,5	44,2	52,5	33,6	42,4
942,42	56,4	56,5	56,2	58,7	41,0	55,8	36,0	43,2
944,2	66,8	68,7	68,4	70,9	50,1	64,9	45,5	54,7
945,98	62,4	65,4	63,7	67,9	49,4	60,0	42,8	51,3
947,69	62,9	65,9	66,9	67,9	48,2	63,0	44,3	53,2
949,49	67,0	68,9	68,8	70,4	50,2	65,1	45,3	54,4
951,2	94,2	98,5	96,3	99,2	85,0	90,3	74,0	85,1

Видно, что информативность коэффициента поглощения на частоте 931,01 см$^{-1}$ минимальна, на частоте 951,2 см$^{-1}$ – максимальна.

Результаты интегральной оценки интенсивности газовыделений колоний бактерий M. Smegmatis при удалении из вектора состояния коэффициентов поглощения, соответствующих указанным частотам, приведены на рис. 2. Кривая 1 соответствует расчету интегральной оценки по всем 12 показателям, кривые 2 и 3 – при удалении наименее информативного и наиболее информативного коэффициентов поглощения, соответственно.

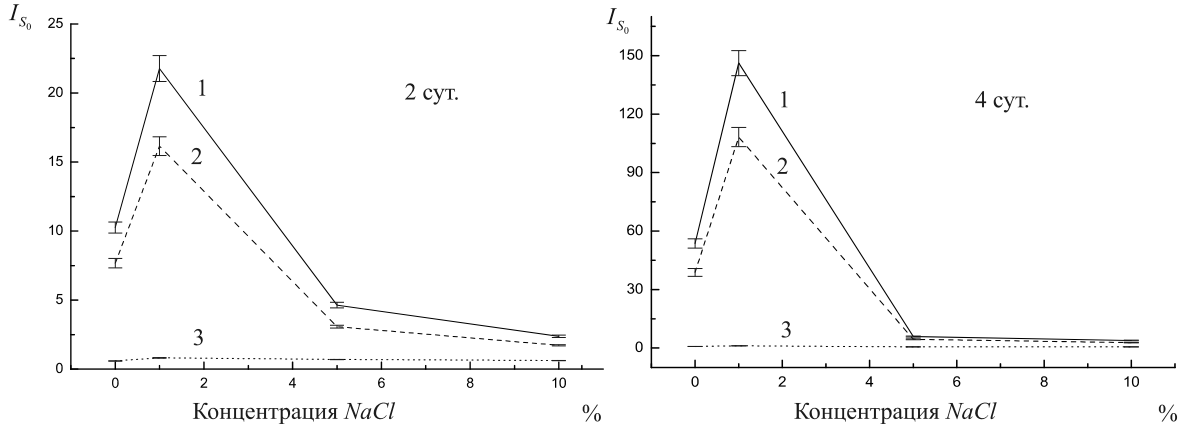

Рис. 2. Зависимость величины интегральной оценки I_{S_0} от концентрации NaCl в различные сроки культивирования колонии бактерий

Из представленных данных следует, что добавление NaCl в концентрации 1% существенно стимулирует интенсивность газовыделений микробактерий на ранней стадии роста колонии, а добавление более высоких концентраций – угнетает.

Данная зависимость сохраняется при меньшем числе экспериментальных показателей, полученных на основе критерия оптимизации (4), (5) и, наоборот, исчезает при удалении из показателей наиболее информативного коэффициента поглощения. Таким образом, данный пример демонстрирует эффективность использованного критерия оптимизации.

В приведенном примере выявление наиболее информативных коэффициентов поглощения позволит упростить дальнейшую расшифровку спектров (выявление молекулярных составляющих в смеси и оценка их концентрации).

Список литературы

Armitage P., Berry G. Statistical Methods in Medical Research. – 3rd ed. – Oxford: Blackwell Scientific Publication, 1994. – 620 p.

Efron B. The Jackknife, the Bootstrap and Other Resampling Plans. // CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph. 38. – Philadelphia: SIAM, 1982. – 92 p.

Manly B. F. J. Randomization, Bootstrap and Monte Carlo Methods in Biology. – London: Chapman and Hall/CRC, 1997. – 424 p.

Баевский Р. М. Оценка и классификация уровней здоровья с точки зрения теории адаптации // Вестник АМН СССР. – 1989. – № 8. – С. 73–78.

Богомолов А. В., Гридин Л. А., Кукушкин Ю. А., Ушаков И. Б. Диагностика состояния человека: математические подходы. – М.: Медицина, 2003. – 464 с.

Генкин А. А. Новая информационная технология анализа медицинских данных (программный комплекс ОМИС). – СПб.: Политехника, 1999. – 191 с.

Дюк В., Эммануэль В. Информационные технологии в медико-биологических исследованиях. – СПб.: Питер, 2003. – 528 с.

Козинец Г. И., Быкова И. А., Сукиасова Т. Г. Кинетика эритрона // Кинетические аспекты гемопоэза / Под ред. Г. И. Козинца и Е. Д. Гольдберга. – Томск: изд-во Томск. ун-та. – 1982. – С. 79–148.

Конрадов А. А. Статистические подходы к анализу многомерных гетерогенных биологических систем // Радиационная биология, радиоэкология. – 1994. – Т. 34. – Вып. 6. – С. 877–886.

Миронкина Ю. Н., Бобров А. Ф. Информационная технология статистического синтеза критериев и алгоритмов оценки функционального состояния человека в прикладных медико-биологических исследованиях // Информационные технологии. – 1998. – № 3. – С. 41–47.

Муха Ю. П., Скворцов М. Г., Авдеюк О. А., и др. Диагностический комплекс основных жизненно важных функций человека по интегральному параметру на основе нейросетевых технологий // Биомед. радиоэлектрон. – 2001. – № 4. – С. 38–41.

Нисевич Н. И., Марчук Г. И., Зубикова И. И., Погожев И. Б. Математическое моделирование вирусного гепатита. – М.: Наука, 1981. – 352 с.

Новицкий В. В., Степовая Е. А., Гольдберг В. Е., Колосова М. В., Рязанцева Н. В., Корчин В. И. Эритроциты и злокачественные образования. – Томск: STT, 2000. – 288 с.

Подвальный С. Л., Матасов А. С., Бырко И. А. Методы многомерной классификации в задачах медицинской диагностики // Машиностроитель. – 2002. – № 8. – С. 59–61.

Степанов Е. В., Миляев В. А., Селиванов Ю. Г. Лазерная ортомолекулярная медицинская диагностика // Успехи физических наук. 2000. – Т. 170, № 4. – С. 458–462.

Ту Дж., Гонсалес Р. Принципы распознавания образов. – М.: Мир, 1978. – 416 с.

Фокин В. А. Критерий оценки состояния сложных биосистем // Известия Томского политехнического университета. – 2004. – Т. 307, № 5. – С. 136–138.