УДК 519.23:578.087.1

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ДАННЫХ ПРИ ОЦЕНКЕ СОСТОЯНИЯ БИОЛОГИЧЕСКИХ СИСТЕМ

В.А. Фокин

ГОУ ВПО «Сибирский государственный медицинский университет», г. Томск
E-mail: fokin@ssmu.tomsk.ru

Методом статистического моделирования данных показано, что использование малых по объему выборок, характеризующих референтное состояние биологических систем, приводит к завышению степени тяжести интегральной оценки состояния. Установлено, что в качестве интегральной оценки состояния биологических систем эффективным является использование асимптотических интегральных оценок, получаемых при неограниченном увеличении объема модельной референтной выборки.

Введение

Анализ медико-биологических данных, разработка методов извлечения из них информации, формирования интегральных оценок состояния биосистем представляют собой активно развивающиеся направления современных информационных технологий в медицинской науке и практике здравоохранения [1–4]. В математической формулировке задача сводится к построению алгоритма и функционального отображения пространства признаков, характеризующих биосистему в одномерное пространство оценок состояний этой системы, определяемых величиной заданного интегрального критерия.

Построение интегральных критериев оценки состояния может приводить к получению эффективных оценок, однако их использование предполагает накопление достаточно больших объемов референтных данных, что не всегда может быть реализовано в условиях отдельного экспериментального исследования. Поэтому статистическое моделирование данных можно рассматривать, как эффективный метод исследования свойств биосистем, результаты которого позволяют, с одной стороны, численно оценивать статистические свойства самого критерия, а с другой – позволяют определять условия, накладываемые на объемы выборок, необходимые для получения устойчивых обобщенных оценок состояния.

Вид критерия оценки состояния

Оценку состояния системы S будем производить по отношению к некоторому референтному состоянию S_0 данной системы. В качестве такого состояния может быть выбрано, например, состояние системы соответствующее здоровому организму. Пусть S_0 и S заданные референтное и оцениваемое состояния, характеризуемые множествами объектов $\{\vec{b}_i | i \in N_{S_0}\}$ и $\{\vec{b}_j | j \in N_S\}$ соответственно. Здесь N_{S_0} и N_S – объемы выборок. Величина количественной оценки состояния некоторого объекта $\vec{b}_i \in S$ может быть охарактеризована его мерой близости к референтному состоянию S_0, при выборе которой следует учитывать конфигурацию области в пространстве признаков, занимаемой референтным состоянием, расположением объектов $\vec{b}_j$ относительно данной области, а также взаимным расположением объектов, представляющих референтное состояние системы. С учетом этих условий, критерий интегральной оценки близости объекта $\vec{b}_i$ к состоянию S_0 можно задать следующим образом [5]:

$$I_{S_0}(\vec{b}_i) = \frac{d(\vec{b}_i, S_0)}{D_{S_0}}, \qquad (*)$$

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

Нормировка на величину D_{S_0} в выражении (*), позволяет учесть вклад в получаемую оценку, как конфигурации области S_0, так и взаимного расположения объектов в ней. Меру компактности D_{S_0} референтного состояния S_0 зададим в следующем виде:

$$D_{S_0} = \frac{1}{N_{S_0}} \sum_{k=1}^{N_{S_0}} \frac{1}{N_{S_0} - 1} \sum_{j=1}^{N_{S_0}-1} d(\vec{b}_k, \vec{b}_j),$$

т. е., как усредненное значение средних расстояний от каждого объекта, относящегося к состоянию S_0, до всех оставшихся. Определенная таким образом величина D_{S_0} представляет собой внутимножественное расстояние [6], конкретный вид которого определяется способом задания расстояния в пространстве признаков. В качестве меры близости объектов в пространстве признаков в биомедицинских задачах эффективно использование расстояния Махаланобиса [7], поскольку при этом естественным образом учитывается взаимозависимость признаков, характеризующих изучаемые биообъекты. Расстояние Махаланобиса между k-м и i-м объектами определяется следующим образом:

$$d_M(\vec{b}_k, \vec{b}_i) = (\vec{b}_k - \vec{b}_i)^T \mathbf{C}_0^{-1} (\vec{b}_k - \vec{b}_i).$$

Здесь $\mathbf{C}_0$ – матрица ковариации признаков, характеризующих состояние S_0. Количественная оценка меры компактности области, характеризующей состояние S_0 в метрике Махаланобиса, будет равна удвоенной размерности пространства признаков [5]:

$$D^*_{S_0} = D_{S_0} = 2m,$$

а выражение для интегральной оценки близости объекта $\vec{b}_i$ к состоянию S_0 примет вид:

$$I_{S_0}(\vec{b}_i) = \frac{1}{2m} d(\vec{b}_i, S_0),$$

где $d(\vec{b}_i, S_0)$ рассчитывается как усредненное расстояние Махаланобиса от объекта $\vec{b}_i$ до S_0,

$$d(\vec{b}_i, S_0) = \frac{1}{N_{S_0}} \sum_{j=1}^{N_{S_0}} d_M(\vec{b}_i, \vec{b}_j),$$

вычисляемое с использованием матрицы ковариации, соответствующей референтному состоянию S_0.

Статистическое моделирование

Основная проблема при использовании критериев, основанных на многомерных методах анализа данных, обусловлена малыми объемами выборок, характеризующих референтное состояние, что приводит к значительной вариабельности оценок, получаемых с их использованием. В этом отражается специфика биомедицинских данных и прежде всего их широкая внутри- и междуиндивидуальная вариабельность, следствием которой является тот факт, что проведение повторных измерений на одной и той же выборке может приводить к различным количественным значениям оцениваемых характеристик. Оценка статистических свойств предложенного выше интегрального критерия (*), представляет собой нетривиальную задачу, решение которой с использованием только аналитических подходов обусловлено значительными трудностями, а в ряде практических случаев невозможно.

В этом случае исследование статистических свойств интегрального критерия может быть эффективно реализовано методами статистического моделирования, результаты которого позволяют, с одной стороны, численно оценивать статистические свойства критерия, а с другой – позволяют определить условия, накладываемые на условия формирования референтных выборок, необходимые для получения устойчивых оценок.

Оценка статистических характеристик интегрального показателя проводилась в два этапа. На первом этапе моделировалось M выборочных множеств X_k ($k=\overline{1,M}$) заданного объема, соответствующих статистическим характеристикам референтного состояния S_0, представленного некоторым выборочным множеством объектов $X:\{\vec{b}_i | i \in \overline{1,N_{S_0}}\}$. Полученные последовательности значений имитируют взятие выборок из одной и той же совокупности и, следовательно, будут свободны от погрешностей, обусловленных влиянием внутри- и междуиндивидуальной вариабельности биологических данных. Далее для каждого множества X_k вычислялись величины оценок $I_{S_0,k}(\vec{b})$, распределение которых в дальнейшем использовалось для исследования статистических свойств интегрального критерия. Здесь вектор $\vec{b}$ характеризует объект, для которого производится оценка. В частности, в качестве его можно рассматривать вектор, соответствующий эталонному представителю состояния S, например, вектор, соответствующий центру класса.

На втором этапе исследовалось, как на величину оценок будут сказываться такие факторы, как объем выборки, соотношение между объемом выборки и количеством совместно анализируемых показателей и т. п. В зависимости от того, известен или нет закон распределения многомерных данных, для статистического моделирования наборов их значений могут применяться различные методы [8–10]. Статистические свойства интегрального критерия оценивались путем вычисления среднего значения интегрального показателя

$$\hat{I}_{S_0}(\vec{b}) = \frac{1}{M} \sum_{k=1}^{M} I_{S_0,k}(\vec{b}),$$

среднего квадратичного отклонения

$$\hat{\sigma}_I(\vec{b}) = \frac{1}{M} \sum_{k=1}^{M} (I_{S_0,k}(\vec{b}) - \hat{I}_{S_0}(\vec{b}))^2.$$

Для оценки вариабельности интегрального показателя рассчитывался коэффициент вариации

$$V = \frac{\hat{\sigma}_I(\vec{b})}{\hat{I}_{S_0}(\vec{b})} \cdot 100\ \%$$

и $(1-p)\cdot 100$ %-го доверительный интервал, как интервал, содержащий значения I_{S_0}, находящиеся между $p/2\cdot 100$ % и $(1-p/2)\cdot 100$ % числом всех значений интегрального показателя в ранжированном ряду оценок. Здесь p – соответствующий уровень статистической значимости. Такой непараметрический способ оценки доверительного интервала позволяет оценивать его без каких-либо предположений относительно вида закона распределения и статистических свойств интегрального показателя.

Результаты моделирования

Исходными данными для формирования модельных выборок и проведения статистических оценок предлагаемого интегрального критерия послужили данные сканирующей электронной микроскопии (СЭМ) поверхностной архитектоники клеток красной крови, полученные коллективом авторов [11–13] по результатам обследований больных при некоторых локализациях онкологических заболеваний II–III стадий, а также здоровых лиц. Поскольку форма эритроцитов и их способность к деформации является следствием комплекса нарушений свойств, организации и метаболизма отдельных компонентов эритроцитов, обусловленных наличием соответствующего патологического процесса, то данные СЭМ могут быть использованы для интегральной оценки степени изменений, происходящих в системе красной крови по выбранному комплексу показателей.

Статистическое оценивание I_{S_0} проводилось с использованием разработанной компьютерной программы [14], путем моделирования выборок, соответствующих объемам N_{S_0}, равным 50, 100, 200, 400, 600, 800 и 1000 наблюдений, в предположении, что данные референтной выборки удовлетворяют многомерному нормальному закону распределе-

ния. Каждая выборка моделировалась от 100 до 1000 раз с шагом 100, по которым в дальнейшем рассчитывались статистические оценки вариабельности величины критерия. Результаты моделирования статистических характеристик интегрального показателя I_{S_0} и его вариабельности для оценки состояния системы красной крови по данным СЭМ при различных локализациях рака для некоторых значений N_{S_0} и M представлены на рис. 1 и 2.

Из анализа полученных результатов следует, что на получение устойчивых оценок существенное значение будет оказывать величина объема референтной выборки N_{S_0}. В частности, при малых объемах выборок наблюдается широкая вариабельность величины I_{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. На рис. 2 приведена зависимость рассчитанных средних значений величины интегрального критерия I_{S_0} от объема модельной выборки для онкологических заболеваний различных локализаций, соответствующих количеству модельных выборок M=500.

Для других объемов модельных выборок зависимости имеют аналогичный вид. Отрезками указаны соответствующие 95 % доверительные интервалы. Интересным результатом статистического моделирования явилось то, что величина оценки зависит от объема референтной выборки, причем малые объемы выборок будут приводить к завы-

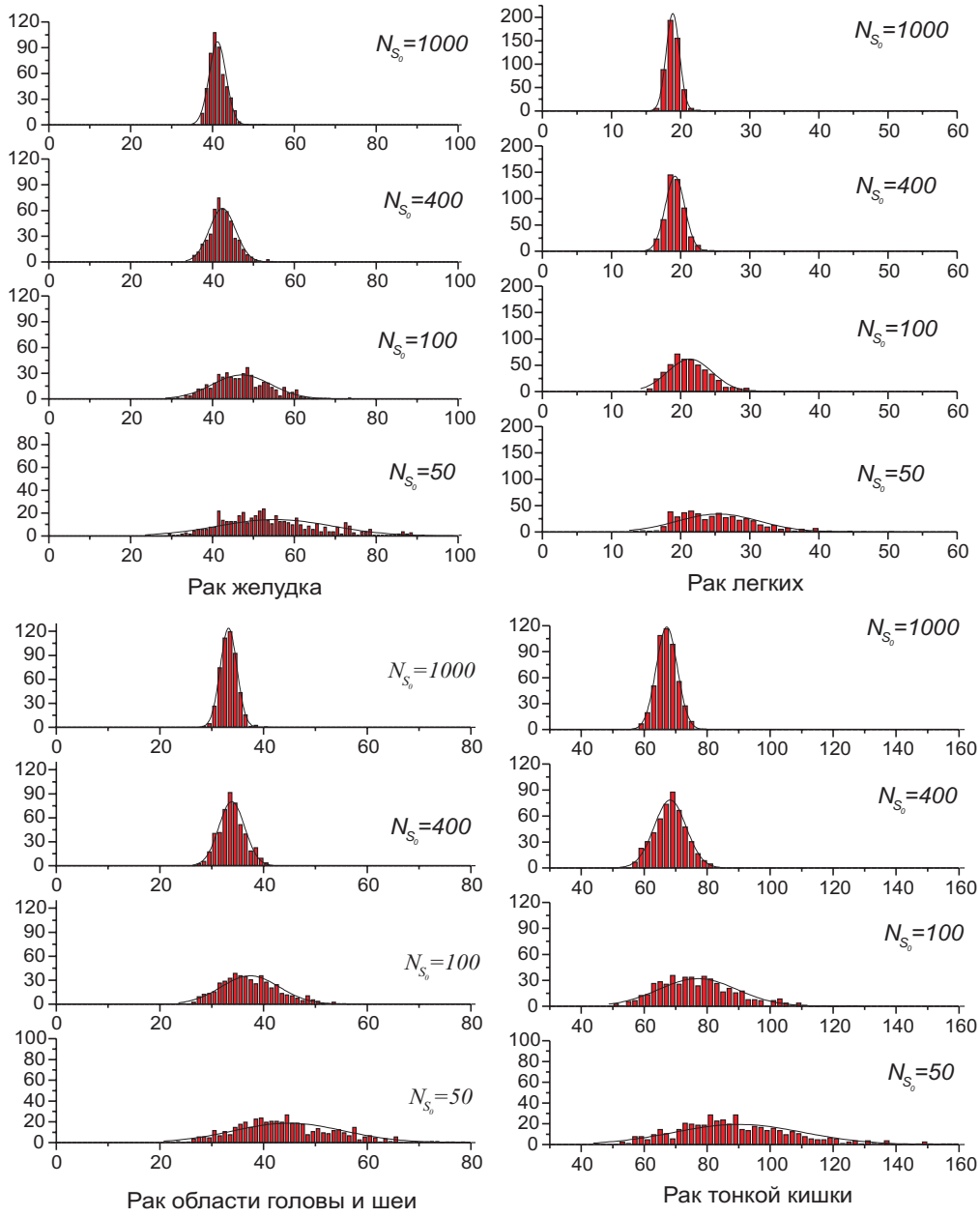

Рис. 1. *Гистограммы частот I_{S_0} для различных объемов N_{S_0} референтного состояния. По оси абсцисс – значение I_{S_0}, по оси ординат – значение частоты. Кривая – аппроксимация нормальным распределением*

шенным оценкам средней величины интегрального показателя. Поэтому может представлять интерес рассмотрение асимптотических оценок, получаемых при неограниченном увеличении объема моделируемой референтной выборки.

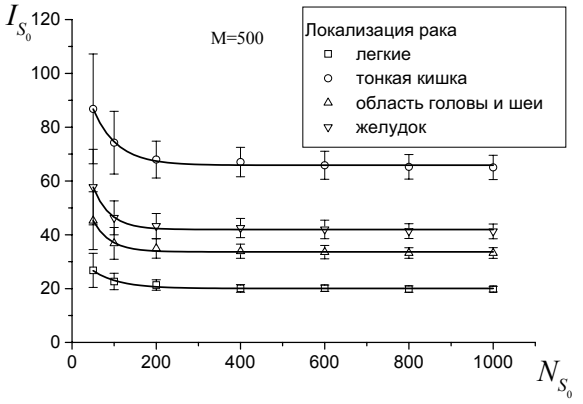

Рис. 2. *Зависимость величины интегрального критерия I_{S_0} от объема модельной выборки N_{S_0} референтного состояния*

Выводы

1. Статистическое моделирование данных референтного состояния биологических систем является эффективным приемом формирования выборок многомерных данных, позволяющим уменьшить вариабельность интегральной оценки состояния исследуемой системы в условиях малых объемов исходных экспериментальных данных.
2. Объем референтной выборки является существенным параметром для количественной оценки состояния биологических систем, причем малые объемы референтных выборок, наряду с большой вариабельностью получаемых интегральных оценок, приводят также к завышению средней величины получаемой интегральной оценки.
3. В качестве интегральной оценки состояния биологических систем эффективно использование асимптотических интегральных оценок, получаемых при неограниченном увеличении объема модельной референтной выборки.

СПИСОК ЛИТЕРАТУРЫ

1. Богомолов А.В., Гридин Л.А., Кукушкин Ю.А., Ушаков И.Б. Диагностика состояния человека: математические подходы. – М.: Медицина, 2003. – 464 с.
2. Генкин А.А. Новая информационная технология анализа медицинских данных (программный комплекс ОМИС). – СПб.: Политехника, 1999. – 191 с.
3. Дюк В., Эммануэль В. Информационные технологии в медико-биологических исследованиях. – СПб.: Питер, 2003. – 528 с.
4. Armitage P., Berry G. Statistical Methods in Medical Research. – 3rd ed. – Oxford: Blackwell Scientific Publication, 1994. – 620 p.
5. Фокин В.А. Критерий оценки состояния сложных биосистем // Известия Томского политехнического университета. – 2004. – Т. 307. – № 5. – С. 136–138.
6. Ту Дж., Гонсалес Р. Принципы распознавания образов. – М.: Мир, 1978. – 416 с.
7. Конрадов А.А. Статистические подходы к анализу многомерных гетерогенных биологических систем // Радиационная биология, радиоэкология. – 1994. – Т. 34. – Вып. 6. – С. 877–886.
8. Ермаков С.М., Михайлов Г.А. Статистическое моделирование. 2-е изд. – М.: Наука, 1982. – 296 с.
9. 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.
10. Manly B.F.J. Randomization, Bootstrap and Monte Carlo Methods in Biology. – London: Chapman and Hall/CRC, 1997. – 424 p.
11. Новицкий В.В., Рязанцева Н.В., Степовая Е.А., Быстрицкий Л.Д., Ткаченко С.Б. Атлас. Клинический патоморфоз эритроцитов. – М.: ГЭОТАР-МЕД, 2003. – 208 с.
12. Новицкий В.В., Степовая Е.А., Гольдберг В.Е., Колосова М.В., Корешкова К.Г., Соколова И.Б., Булавина Я.В. Обратимая агрегация и поверхностная архитектоника эритроцитов периферической крови у больных раком легкого до и в ходе проведения противоопухолевой полихимиотерапии // Экспериментальная и клиническая фармакология. – 1999. – Т. 62. – № 5. – С. 28–30.
13. 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.
14. Свид. № 2006614010 РФ. Программа для ЭВМ «StatSys» / В.А. Фокин, И.С. Хакимов, О.Ю. Никифорова; Заявка № 2006613281; Заявлено 29.09.2006; Опубл. 22.11.2006.

Поступила 04.10.2007 г.