УДК 519.254

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ ВЫЯВЛЕНИЯ ОСНОВНЫХ ЗАКОНОМЕРНОСТЕЙ ВРЕМЕННОГО ИЗМЕНЕНИЯ ПОКАЗАТЕЛЕЙ БИОСИСТЕМЫ

[1]Гергет О.М., [1]Милешин А.А.

[1]Национальный исследовательский Томский политехнический университет, Томск, Россия (634050, г. Томск, пр. Ленина, 30), e-mail: Olgagerget@mail.ru

Статья является результатом научных исследований сотрудников кафедры Прикладной математики Томского политехнического университета, работающих в коллективе научной школы «Разработка физических основ программного обеспечения энерго-информационного представления функциональных особенностей организма в задачах лечебно-профилактической медицины» и посвящена разработке информационной медицинской системы и применению математических методов для выявления закономерностей временного изменения показателей биохимии крови на основе статистического анализа. В статье приведена структура информационной системы, которая включает сервисы: восстановления пропусков в данных; выявления наличия сезонных ритмов; выделения трендов во временных рядах; определения сезонной декомпозиции. Изложены программно реализованные основные математические методы. Приведены результаты исследования. Сформирован стандарт поведения показателей биохимии крови во времени. Проведена оценка состояния здоровья организма человека.

Ключевые слова: моделирование, тренд, сезонные колебания, аномальные наблюдения, медицина.

INFORMATION TECHNOLOGIES OF BIOSYSTEM INDEXES TIME CHANGE MAIN TENDENCIES DETERMINATION

[1]Gerget O.M., Mileshin A.A.

[1]National Research Tomsk Polytechnic University, Tomsk, Russia, 634050, Tomsk, Lenin Avenue, 30

The article includes results of scientific results achieved at department of Applied Mathematics at Tomsk Polytechnic University. Investigators were working in team of scientific school "Developing principles of software providing energy-information organism functional characteristics representation within the context of preventive and curative medicine" and is devoted to developing of information medical system and application of mathematical methods for determination time change tendencies of blood biochemistry indexes, based on statistival analysis. The article brings information system structure which includes services: recovery of omissions in data; determination seasonal rhythms' existence; determination of trends in time series; seasonal decomposition. Main mathematical methods that are realized in program are stated. The article reveals results of research. Standart of blood biochemistry indexes' behavior in time is formed. Estimation of patient organism health state and efficiency of provided treatment is done.

Keywords: modeling, trend, seasonal rhythms, anomalistic observations, medicine.

Введение

В настоящее время биология и медицина стремительно отходят от вербального описания и основываются на математических моделях и информационных технологиях Успешное решение биомедицинских задач невозможно без создания соответствующих информационных систем. Одним из наиболее сложных и трудоемких процессов проектирования информационной системы является выявление закономерностей из имеющихся массивов данных. Он не всегда заканчивается успешно, поскольку базы данных содержат разнотипную, противоречивую и неполную информацию. Большинство существующих в настоящее время информационных технологий ориентированы на решение конкретных практических задач и являются узконаправленными, сложными,

дорогостоящими, что делает их непригодными для массового применения в медицинских учреждениях. В связи с этим авторами разработана система, которая позволяет выявить закономерности временного изменения показателей биосистемы и включает такие важные сервисы, как восстановление пропусков в данных, выявление наличия сезонных ритмов, выделение трендов во временных рядах, определение сезонной декомпозиции.

Целью исследования является выявление закономерности временного изменения показателей биосистемы на основе статистического анализа.

Структура информационной медицинской системы

Для осуществления поставленной цели в информационной медицинской системе разработаны сервисы с применением распараллеленных вычислений. Среди них:

1. Сервис восстановления пропусков в исходных данных

Пусть значения показателей известны в моменты времени t_j, j=1,…,n. Построим на временном интервале $[t_1, t_n]$ функцию $S_y(t)$, интерполирующую y=f(t) так, что на каждом произвольном отрезке $[t_j, t_{j+1}]$, лежащем внутри интервала $[t_1, t_n]$, функция $S_y(t)$ являлась полиномом, а в узлах имела непрерывные производные.

В качестве $S_y(t)$ выбран сплайн третьей степени, который в узлах t_j имеет непрерывные 1-ю, и 2-ю производные, и на каждом из отрезков $[t_j, t_{j+1}]$ принимает вид:

$$S_y(t) = y_j + b_j(t - t_j) + c_j(t - t_j)^2 + d_j(t - t_j)^3,$$

где y_j – значение показателей в момент t_j

b_j, c_j, d_j – коэффициенты, подлежащие определению.

Требования непрерывности функции $S_y(t)$, ее 1-й и 2-й производных, дает 3(n–2) условий для определения коэффициентов. Условия интерполирования в точке t_n, приводят к соотношениям для вычисления коэффициентов b_j, c_j, d_j на разных интервалах аппроксимации:

$$d_j = \frac{c_{j+1} - c_j}{3\Delta t_j},$$

где $\Delta t_j = t_{j+1} - t_j$

$$b_j = \frac{y_{j+1} - y_j}{\Delta t_j} - \frac{\Delta t_j}{3}(2c_j + c_j + 1), \quad j = 1, \ldots, n-1$$

$$\beta_j c_{j-1} + 2c_j + \lambda_j c_{j+1} = \frac{3}{\Delta t_j + \Delta t_{j-1}}\left(\frac{y_{j+1} - y_j}{\Delta t_j} - \frac{y_j - y_{j-1}}{\Delta t_{j-1}}\right), \quad j = 2, \ldots, n-1$$

где $\lambda_j = \Delta t_j(\Delta t_{j-1} + \Delta t_j)^{-1}$;

$$\beta_j = 1 - \lambda_j.$$

Из краевых условий запишем уравнения следующего вида:

$$S'_y(t_1) = y'_1;\ S'_y(t_n) = y'_n,$$

тогда $2c_1 + c_2 = \frac{3}{\Delta t_1}\left(\frac{y_2 - y_1}{\Delta t_1} - y_1'\right)$,

$c_{n-1} + 2c_n = \frac{3}{\Delta t_{n-1}}\left(y_n' - \frac{y_n - y_{n-1}}{\Delta t_{n-1}}\right)$;

$S_y'''(t_1) = 6\Delta_1^{(3)}; S_y'''(t_n) = 6\Delta_{n-3}^{(3)}$,

где $\Delta_1^{(3)}, \Delta_{n-3}^{(3)}$ – третьи разделенные разности от функции $y = f(t)$ по точкам t_1, t_2, t_3, t_4, t_{n-3},t_{n-2}, t_{n-1}, t_n соответственно.

Тогда $-c_1 + c_2 = \Delta t_1 \Delta_1^{(3)}$,

$c_{n-1} - c_n = -\Delta t_{n-1}\Delta_{n-3}^{(3)}$.

Приведенные системы уравнений решаются методом Гаусса.

Оценка качества восстановленного показателя осуществлялась с помощью нахождения коэффициента расхождения, предложенного Тейлором:

$$B = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_k - y_k)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} y_k^2}} = \frac{\sqrt{\sum_{i=1}^{N}(\hat{y}_k - y_k)^2}}{\sqrt{\sum_{i=1}^{N} y_k^2}}$$

где $\hat{y}_k$ – предсказанное значение для y_k ; y_k – фактическое значение.

2. Сервис выявления тренд-циклической компоненты

Оценка наличия сезонных ритмов во временных рядах осуществлялась на основе функции автокорреляции и её графического представления – коррелограммы. При помощи анализа коррелограммы можно выявить структуру ряда. Если наиболее высоким оказался коэффициент корреляции первого порядка, то исследуемый ряд содержит только тенденцию, если коэффициент автокорреляции порядка h, то ряд содержит циклические колебания с периодичностью в h моментов времени [3].

Последовательность коэффициентов автокорреляции со смещениями 1, 2, 3 и т.д. называют автокорреляционной функцией, значения которой находятся в диапазоне [-1; 1].

Автокорреляционную функцию целесообразно использовать для выделения во временном ряде наличия трендовой и сезонной компонент.

3. Сервис выделения трендов во временных рядах

Выявление наличия неслучайной составляющей сводилось к проверке гипотезы о неизменности среднего значения временного ряда с использованием критерия серий. При его использовании определяется медиана временного ряда, и образуются «серии» из плюсов и минусов по следующему правилу:

$$y_t = \begin{cases} +, & y_t > y_{med}, \\ -, & y_t < y_{med} \end{cases}$$

Элементы временного ряда, равные y_{med}, в полученной таким образом последовательности не учитываются. Под «серией» понимается последовательность подряд идущих плюсов или подряд идущих минусов. Наличие неслучайной составляющей во временном ряде определяется из условия:

$\begin{cases} v(n) > \left[\frac{1}{2}(n+2-1{,}96\sqrt{n-1})\right], \\ K_{\max} < [3{,}3(\lg n+1)] \end{cases}$ где $v(n)$ – общее число серий, K_{max} – длина наибольшей серии, [ ] – целая часть от числа [1].

Для построения тренда использовались два метода: скользящих средних и экспоненциального сглаживания [4].

Метод скользящих средних заключается в следующем: 1) определяем количество наблюдений, входящих в интервал сглаживания; 2) вычисляем среднее значение наблюдений в интервале сглаживания по формуле:

$$\bar{y}_t = \frac{1}{m} \sum_{i=t-\frac{m-1}{2}}^{t+\frac{m-1}{2}} y_i ,$$

где m – количество наблюдений, входящих в интервал сглаживания. И так до тех пор, пока в интервал сглаживания не войдет последнее значение временного ряда.

Альтернативный подход к устранению колебаний в ряде значений состоит в использовании метода экспоненциального сглаживания. Каждое сглаженное значение рассчитывается путем сочетания предыдущего сглаженного значения и текущего значения временного ряда. В этом случае текущее значение временного ряда взвешивается с учётом сглаживающей константы:

$$S_t = \alpha y_t - (1-\alpha) S_{t-1},$$

где S_t – текущее сглаженное значение;

y_t – текущее значение временного ряда;

S_{t-1} – предыдущее сглаженное значение;

α – сглаживающая константа, значение которой варьируется в диапазоне от 0 до 1 [5].

4. Сервис оценки сезонной декомпозиции

Для определения сезонной составляющей разработан алгоритм сезонной декомпозиции.

1. Выделение тренда (метод скользящих средних).
2. Формирование сезонной компоненты (разность между исходным и сглаженным рядом).
3. Вычисление сезонной компоненты (среднее всех значений ряда, соответствующих данной точке сезонного интервала).

4. Определение случайной составляющей.

Информационная система, в состав которой включены данные сервисы, позволяет осуществить комплексный подход к диагностике и прогнозированию состояния здоровья организма человека посредством объединения в единое целое процессов анализа и контроля информации и организации оперативного обмена данными в едином информационном пространстве. Параллельный режим обработки данных обеспечивает высокую загрузку вычислительных ресурсов посредством распределения одной сложной задачи на несколько вычислительных узлов.

Результаты исследования

Экспериментальная выборка составляла 527 объектов исследования. Каждый объект описан вектором состояния $x = (x_1, \dots, x_{23})$. Исследование проводилось в динамике 23 измерения с периодичностью 1 неделя.

Для успешного решения задачи выявления закономерностей временного изменения показателей биосистемы необходимо выявить и удалить аномальные наблюдения в данных. С этой целью был использован метод Ирвина [2]:

$$\lambda_t = \frac{|y_t - y_{t-1}|}{S_y},$$

где $$S_y = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(y_t - \bar{y})^2},$$

$$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t.$$

Для проверки гипотезы о наличии аномальных наблюдений во временном ряде был задан уровень значимости равный 0,05. Если полученное λ_t превышает табличное значение, то элемент y_t считается аномальным наблюдением и заменяется на расчетное значение (среднее из двух соседних значений).

Полученные после обнаружения аномальных значений временные ряды признаков были проверены на наличие тренда с помощью критерия серий. Как показали исследования, во всех исследованных временных рядах присутствует тренд. Так, например, для показателя «Сосудистый эндотелиальный фактор роста» (VEGF) были получены следующие значения: медиана равна 12,83, общее количество серий – 3, максимальная длина серии – 5.

При n = 23

$$\left[\frac{1}{2}(n + 2 - 1{,}96\sqrt{n-1}\right] = 7$$

$$[3{,}3(lgn + 1] = 6.$$

В данном примере оба неравенства из условия наличия неслучайной составляющей нарушены, что свидетельствует о присутствии во временном ряде неслучайной составляющей.

Для проверки временных рядов на наличие сезонной составляющей использовалась автокорреляционная функция и её графическое представление – коррелограмма. Коррелограмма для признака FEGF представлена на рисунке 1.

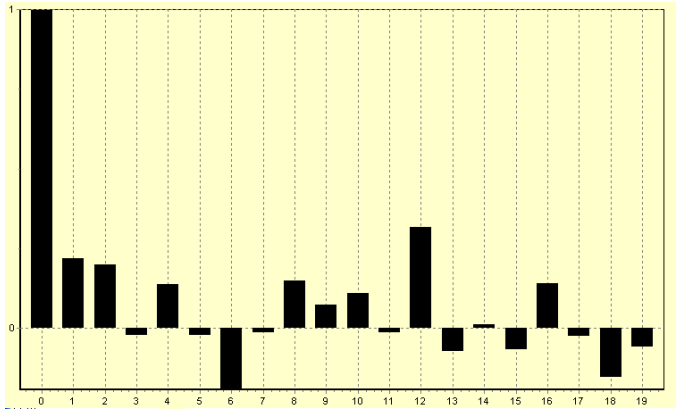

Рис. 1. Показатель VEGF. Период сезонной составляющей – 12

Тренды были построены для тех временных рядов, в которых присутствует сезонная составляющая (по результатам проверки с использованием коррелограммы). В случае, когда для построения тренда был использован метод скользящего среднего, интервал сглаживания для каждого временного ряда был выбран равным периоду сезонных колебаний. При использовании метода экспоненциального сглаживания константа была выбрана равной 0.3. Построенные тренды для признака VEGF представлены на рисунке 2.

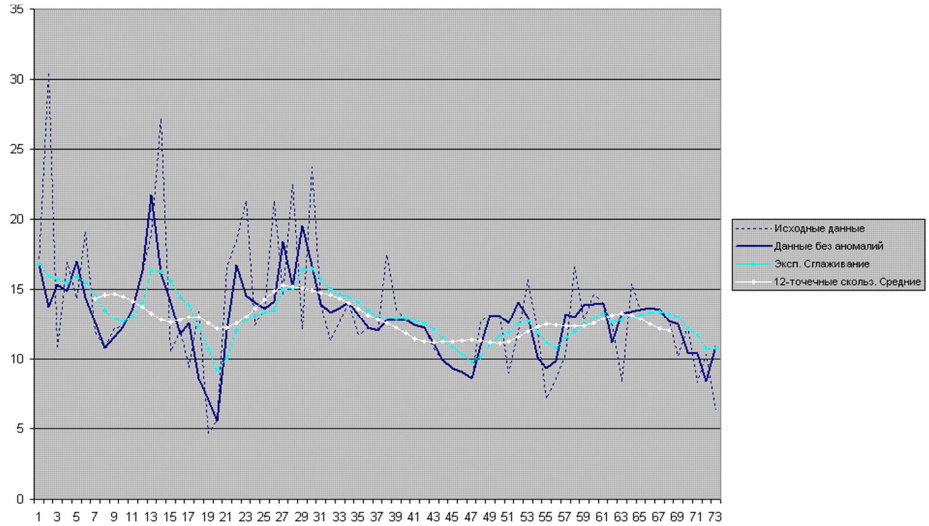

Рис. 2. Тренды для признака VEGF

На основе анализа графического представления временных рядов и их трендов для декомпозиции была выбрана аддитивная модель: X = TC +S + I, где TC – тренд-циклическая компонента, S – сезонная компонента, I – случайная компонента.

Тренд-циклическая компонента получена с помощью метода скользящих средних. Найдены разности между значениями исходного временного ряда и выделенной тренд-циклической компоненты. Вычислена сезонная компонента, как среднее всех значений ряда, соответствующих данной точке сезонного интервала. Получена случайная компонента, как разность значений исходного временного ряда и суммы значений тренд-циклической компоненты и сезонной компоненты. График декомпозиции представлен на рисунке 3.

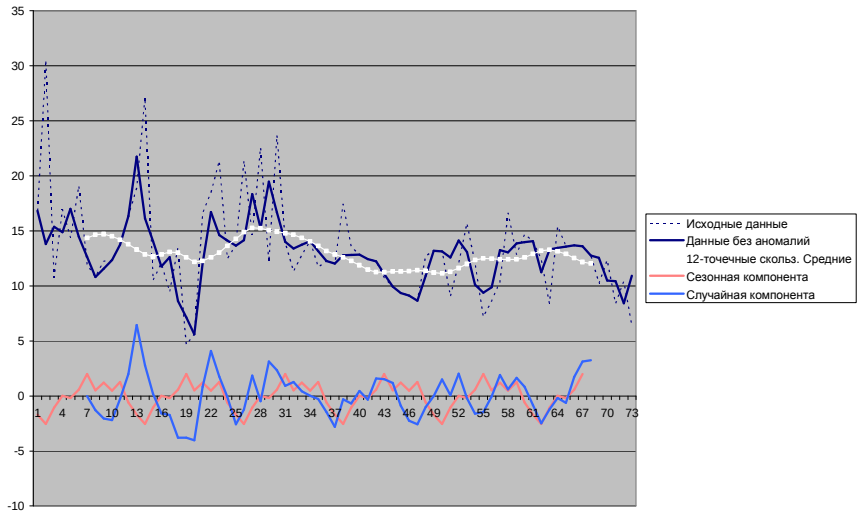

Рис. 3. Декомпозиция временного ряда для признака VEGF

Заключение

Рассмотренные в статье алгоритмы и методы направлены на решение одной из важных проблем: создания эффективных инструментов решения задачи диагностики и прогнозирования состояния здоровья людей. Анализ исследованных временных рядов позволяет представить поведение показателей биохимии крови у здоровых людей.

Апробация информационной медицинской системы на реальных данных показала, что качество решения по вышеизложенному алгоритму удовлетворяет требованиям практического врача. Данная система позволяет выделить данные, в которых присутствует сезонная составляющая, сформировать стандарт поведения исследованных показателей во времени, а также на основе результатов исследования дает возможность оперативно оценить состояние здоровья человека.

Дальнейшие исследования связаны с разработкой магистральных технологий для выявления закономерности реакции организма на условия жизнедеятельности.

Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-90902 мол_ин_нр.

Список литературы

1. Орлова И.В., Половников В.А. Экономико-математические методы и модели: компьютерное моделирование : учеб. пособие. – М. : Вузовский учебник, 2007. – 365 с.
2. Ричард Томас Количественный анализ хозяйственных операций и управленческих решений. – М. : Дело и Сервис, 2003. – 430 с.
3. Box G.E.P., Jenkins G.M. Time Series Analysis: Forecasting and Control. - 2nd ed. - San Francisco : Holden-Day, 1976.
4. Hoel P.G. Elementary statistics. - Second Edition. – Wiley, 1971. – 309 p.
5. Siegel A.F. Practical business-statistics. - 4th edition. - 2004. – 1056 p.

Рецензенты:

Михалев Е.В., д.м.н., профессор, зав. кафедрой педиатрии ФПК и ППС, ГБОУ ВПО «СибГМУ» Минздрава России, г. Томск.

Кочегуров В.А., д.т.н, профессор, ФГБОУ ВПО «Национальный исследовательский Томский политехнический университет», г. Томск.