ПРОГНОЗИРОВАНИЕ НА ОСНОВЕ ЭВОЛЮЦИОННОГО МЕТОДА ОБРАБОТКИ МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ.

Цыганкова И.А.

УДК 004.8.023

Цыганкова И.А. **Прогнозирование на основе эволюционного метода обработки медико-биологических данных.**

Аннотация. В работе представлен метод обработки многомерных плохо формализованных массивов медико-биологической информации, базирующийся на эволюционном подходе к решению экстремальных задач функции многих переменных. Предлагаемый метод позволяет прогнозировать результаты лечения с учетом медико-биологических и социальных особенностей пациентов. Приведены результаты численного эксперимента.

Ключевые слова: обработка данных, эволюционный метод, медико-биологическая информация, прогнозирование

Tsygankova I.A. **Forecasting based on evolutionary method of processing medical and biologic data.**

Abstract: Method of intellectual processing of poorly formalized multivariable diverse arrays of biomedical information, based on evolutional method for solving of extreme tasks of multivariable function, is presented in the article. The proposed method allows predicting treatment results take account of biomedical and social features of the patients. Results of numerical experiment are adduced.

Keywords: data processing, evolutionary method, the medical and biologic information, forecasting

1. Введение. Рост требований к качеству жизни, появление новых диагностических и лечебных технологий привели к значительному увеличению стоимости медицинских услуг. Это резко обострило проблему оптимизации затрат на лечение и профилактику заболеваний, как для отдельных пациентов, так и для медицинских организаций различного уровня. Решение этой проблемы может быть получено только современными методами оптимизации и прогнозирования результатов лечения, учитывающими медико-биологические и социальные особенности пациентов.

Развитие вычислительной техники и информационных технологий позволяет в настоящее время перейти к решению задач прогнозирования в медицине, используя интеллектуальные методы анализа данных [1–3]. Особенностями реальных медико-биологических данных являются: высокая размерность и разнотипность данных, большое количество "шумящих" и дублирующих признаков, пропущенные и аномальные значения. В такой ситуации

Труды СПИИРАН. 2009. Вып. 9. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2009. Issue 9. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

эффективными становятся методы, основанные на эволюционном подходе, которые в отличие от традиционных методов поиска оптимального решения, ориентированы на получение «наилучшего» (приемлемого) решения по сравнению с полученным ранее или заданным в качестве начального.

2. Постановка задачи. Рассматривается задача прогнозирования результатов лечения при заданной тактике лечения на примере кожного хронического заболевания псориазом. Исходная информация о больных представлена в виде числовых таблиц «объект-свойство» с описанием входных и выходных параметров (признаков, характеристик) пациентов.

К входным параметрам относятся индивидуальные сведения о больном: анамнез, сопутствующие заболевания, клинико-функциональные, метаболические и иммунологические показатели, тактика лечения.

Выходными (целевыми) параметрами являются: продолжительность пребывания пациента в стационаре (количество койко-дней), продолжительность лечения до наступления улучшения (эффект лечения), продолжительность периода ремиссии, наличие (или отсутствие) типичных остаточных поражений на коже, число обострений болезни в год.

Входные параметры в различной степени влияют на выходные параметры, но какие из них оказывают наиболее существенное влияние на целевые параметры, и какой моделью описываются зависимости их влияния, неизвестно.

В общем случае исходная информация об объектах представлена в виде матрицы

$$\boldsymbol{Z} = (Z_1, Z_2, \cdots, Z_i, \cdots, Z_N),$$

где $Z_i = \left(z_{i1}, z_{i2}, \cdots, z_{ij}, \cdots, z_{iM}\right)'$ — вектор анализируемых параметров (свойств, признаков) i-го объекта. Каждый параметр z_{ij} принимает значение из множества допустимых значений. Вся совокупность параметров объектов делится на входные параметры $V = (v_1, v_2, \cdots, v_t)$ и выходные параметры $Y = (y_1, y_2, \cdots, y_s)$. Входные параметры V являются разнотипными, измеряются в количественных и качественных шкалах.

Труды СПИИРАН. 2009. Вып. 9. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2009. Issue 9. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Обозначим через $X = (x_1, x_2, \cdots, x_m)$ параметры, значения которых измеряются в количественных шкалах, а через $U = (u_1, u_2, \cdots, u_h)$ — параметры, значения которых измеряются в качественных (номинальных и порядковых) шкалах. Вектор выходных параметров Y для сформулированной задачи измеряется в количественной шкале.

Требуется с приемлемой точностью предсказать значения неизвестных выходных параметров нового объекта по его известным входным параметрам.

Рассматриваемая задача прогнозирования является плохо формализованной в силу того, что вся информация об объектах представлена лишь набором параметров, о которых нельзя сколько-нибудь определенно сказать, что они полны, непротиворечивы и неискаженны. При таких исходных данных будем использовать модель «черного ящика», а при построении алгоритмов анализа данных - опираться только на массивы прецедентов и гипотезу о монотонности пространства решений: похожие входные ситуации приводят к похожим выходным реакциям системы.

3. Эволюционный метод обработки. Решение задачи прогнозирования с помощью предлагаемого метода состоит из нескольких этапов:

- предобработка данных;
- подбор весовых параметров в процессе обучения;
- предсказание значений целевых параметров.

Этап предобработки включает в себя:

- структуризацию данных;
- выявление и устранение аномальных и пропущенных значений;
- кодировку и нормировку данных, измеряемых в непрерывных шкалах.

Параметры, измеряемые в дискретных шкалах и имеющие число градаций больше двух, преобразуются в совокупность бинарных величин.

Введем вектор $G = (g_1, g_2, \cdots, g_j, \cdots, g_\eta)$, где $g_j, (j = 1,2, \cdots, \eta)$ — бинарные признаки объектов. На этапе предобработки все множество исследуемых объектов разбивается на подмножества (выборки) в соответствии со значениями g_j. Общее количество таких выборок составит c^η, где η — количество бинарных величин, c — количество вариантов (альтернатив) группировки объектов по каждому бинарному признаку g_j. Возможны следующие варианты группировки объектов:

Труды СПИИРАН. 2009. Вып. 9. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2009. Issue 9. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

– в выборку попадают объекты вне зависимости от значения признака g_j;
– в выборку попадают объекты, для которых $g_j = 0$;
– в выборку попадают объекты, для которых $g_j = 1$.

Один и тот же объект может оказаться в нескольких выборках, которые имеют различное количество объектов. В дальнейшем используются только информативно значимые выборки, в которых количество объектов значительно больше числа количественных входных параметров.

На следующем этапе (процесс обучения) для каждой информативно значимой выборки определяются веса входных параметров X. Определение весовых коэффициентов базируется на эволюционном подходе к решению экстремальных задач функции многих переменных и методе случайного поиска. Обозначим вектор весов через

$$W = \left(w_1, w_2, \cdots, w_j, \cdots w_m\right),$$

где $w_j, (j = 1,2, \cdots, m)$ — весовые коэффициенты входных параметров.

Каждый O_i объект может быть представлен в виде вектора многомерного пространства R^p количественных параметров

$$O_i = \left\{x_1, x_2, \cdots, x_j, \cdots, x_m, y\right\},$$

где x_j — входные параметры объекта, y — выходной (целевой) параметр объекта, $p = m + 1$ — общее количество параметров многомерного пространства.

В этом случае задача определения искомого параметра y по известным входным параметрам $X = \left(x_1, x_2, \cdots, x_j, \cdots, x_m\right)$ сводится к задаче интерполяции функции $y = f(X)$, заданной в узлах p -мерной нерегулярной сетки.

Так как степень гладкости функции $f(X)$ неизвестна, для ее интерполяции во всей области определения предлагается использовать функцию вида

$$f(X) \approx y_r\left(d(X, W)\right),$$

Труды СПИИРАН. 2009. Вып. 9. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2009. Issue 9. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

где d — мера близости между объектами. В качестве меры близости между объектами i и l рассматривается "взвешенное" евклидово расстояние

$$d_{il} = \sqrt{\sum_{j=1}^{m} w_j \left(x_{ji} - x_{jl}\right)^2}, \ 0 \le w_j \le 1. \tag{1}$$

Подбор значений весовых коэффициентов W проводится с использованием метода Монте-Карло.

Чтобы обеспечить необходимую точность вычисления прогнозируемого параметра, введем критерий, который минимизирует среднюю абсолютную ошибку прогноза

$$Q(w) = \frac{1}{N_g} \sum_{i=1}^{N_g} |y_i - y_{ri}(d)| \to \min, \tag{2}$$

где $|y - y_r(d)|$ — разность между наблюдаемым и расчетным значениями выходного параметра, N_g — объем исследуемой выборки.

В случае если целевая функция представляет собой комплекс выходных параметров, то априори задаются коэффициенты значимости $\vartheta_j, \ (j = 1, 2, \ \cdots, s)$ для каждого прогнозируемого параметра. Значения коэффициентов ϑ_j выбираются из интервала $[0, 1]$, и для них должно выполняться условие нормировки

$$\sum_{j=1}^{s} \vartheta_j = 1,$$

где s — количество прогнозируемых параметров.

Тогда критерий (2) может быть представлен в виде

$$Q(w) = \frac{1}{N_g} \sum_{i=1}^{N_g} \sum_{j=1}^{s} \vartheta_j \left| y_i^{(j)} - y_{ri}^{(j)}(d) \right| \to \min. \tag{3}$$

Для определения расчетных значений y_{ri} задачу многомерной интерполяции функции $y = f(X)$, заданной в узлах нерегулярной сетки, сведем к задаче одномерной экстраполяции функций $y_{ri}(d)$ $(i = 1, 2, \cdots, N_g)$, в окрестностях каждого i -го узла многомерной сетки. Для этого, относительно каждого i -го узла сетки пространства R^p

Труды СПИИРАН. 2009. Вып. 9. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2009. Issue 9. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

по формуле (1) определяются расстояния между ним и остальными узлами, в которых заданы значения функции y. Затем полученные расстояния ранжируются в порядке возрастания. Ранжированный вектор расстояний обозначим

$$D_i = (d_{i1}, d_{i2}, \cdots, d_{il}, \cdots, d_{i(N_g-1)}).$$

Далее, имея массив, состоящий из пар чисел (d_k, y_k) $\left(k = 1,2, \cdots, N_g - 1\right)$ решается задача экстраполяции дискретной зависимости $y(d_k)$ непрерывной функцией $y_r(d)$. При построении приближающей функции $y_r(d)$ используются только n ближайших узлов $(n < N_g - 1)$. В общем случае величина n определяется в процессе предварительного вычислительного эксперимента. В качестве модели для приближения используется квадратичный полином

$$y_r(d) = \sum_{i=0}^{2} a_i d^i,$$

где коэффициенты a_i определяются из условия минимизации функционала

$$\sigma = \sum_{k=1}^{n} [y_k - y_r(d_k, a_i)]^2 \rightarrow \min.$$

Итеративное уточнение критерия $Q(w)$, вычисляемого по формулам (2) или (3), продолжается до тех пор, пока:

– число итераций, на протяжении которых не улучшается решение, станет больше заранее заданного значения;
– либо пока расчетное значение средней абсолютной ошибки прогноза не упадет ниже априори заданной величины допустимой погрешности;
– либо пока не будет превышено максимальное время вычислений.

Следует отметить, что особенностью эволюционного вычислительного процесса является то, что он может быть остановлен и продолжен в любой момент времени.

Следующий этап решения задачи — использование полученных при обучении результатов для прогнозирования искомых целевых параметров нового объекта по его известным входным характеристи-

Труды СПИИРАН. 2009. Вып. 9. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2009. Issue 9. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

кам. Для этого сначала выявляются те информативные выборки, в которые попадает новый объект с учетом своих качественных признаков. Для дальнейшего анализа используется та выборка, в которой ошибка прогноза имеет наименьшее значение. Расчет каждого целевого параметра нового объекта сводится к задаче экстраполяции функции $y_r(d)$ в окрестности узла сетки этого объекта.

В дальнейшем, после того как становятся известны выходные параметры нового объекта, объект пополняет обучающие выборки, и проводится уточнение весовых коэффициентов в соответствии с изложенным методом. Таким образом, прогнозирование целевых параметров является не разовой операцией, а процессом, в ходе которого постоянно выполняются сбор, очистка и консолидация исходных данных, уточнение весовых параметров и верификация результатов.

4. Численный эксперимент. Для оценки эффективности разработанного метода прогнозирования проведен численный эксперимент с использованием реальных медико-биологических данных больных псориазом, полученных в лечебных медицинских учреждениях Санкт-Петербурга. При проведении численного эксперимента использован программный комплекс поддержки принятия врачебных решений, структура которого описана в работе [4]. Объем исходной выборки пациентов составил 308 человек. Из них случайным образом отобраны 45 пациентов, которые составили контрольную выборку. Общее число параметров составило 44, их которых 39 — входные параметры, 5 — выходные параметры. Обобщенные результаты расчетных исследований по оценке прогноза целевых параметров сведены в таблицу.

Результаты прогноза выходных параметров

Прогнозируемый параметр	Средняя ошибка прогноза
Период лечения в стационаре (количество койко-дней)	0,101
Эффект лечения (период острой стадии)	0,112
Число обострений в год	0,139
Степень разрешения (остаточные поражения на коже)	0,163
Период ремиссии	0,167

Труды СПИИРАН. 2009. Вып. 9. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2009. Issue 9. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Как видно из таблицы, средняя абсолютная ошибка прогноза составила 10–17%. Достоверность полученных результатов подтверждена расчетами на контрольной выборке.

5. Заключение. Предлагаемый метод прогнозирования может быть использован в любой предметной области, где сведения об объектах сведены в информационные массивы большого объема, описываются в протоколах «вход-выход», и для них справедлива гипотеза о монотонности принятия решений в локальной области.

Разработанный метод интеллектуальной обработки многомерных разнотипных массивов медико-биологической информации позволяет подобрать весовые коэффициенты входных параметров, не снижая размерности признакового пространства, что в свою очередь позволяет исключить потерю значимой информации и учесть слабые связи в рассматриваемых информационных массивах. Проведенные расчетные исследования оценки прогнозирования целевых параметров показали высокую эффективность предлагаемого метода.

Литература

1. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. Новосибирск: Изд-во Ин-та математики, 1999. 270 с.
2. *Корнеев В.В., Гареев А.Ф., Васютин С.В., Райх В.В.* Базы данных. Интеллектуальная обработка информации. М.: Нолидж, 2001. 496 с.
3. *Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.* Технологии анализа данных: Data Mining, Visual Mining, OLAP. СПб.: БХВ-Петербург, 2007. 275 с.
4. *Цыганкова И.А.* Программный комплекс системы поддержки принятия врачебных решений // Программные продукты и системы. 2008. №4. С.155–158

Цыганкова Ирина Александровна — канд. техн. наук, старший науч. сотр. лаборатории прикладной информатики Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН). Область научных интересов: разработка методов интеллектуального анализа плохо формализованных данных. Число научных публикаций — 58. Адрес: СПИИРАН, 14 линия В.О., д. 39, Санкт-Петербург, 199178, РФ; pallada-ltd@infopro.spb.su; р.т. +7(812)328-1919, м.т. +7(921)376-7269

Tsygankova Irina Alexandrovna — Ph.D., Senior researcher, Laboratory applied informatics, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: Development of methods of the intellectual analysis of badly formalized data. The number of publications — 58. Adress: SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; pallada-ltd@infopro.spb.su ; office phone +7(812)328-1919, fax +7(812)328-4450.

Поддержка исследований. В публикации представлены результаты исследований, поддержанные грантом РФФИ 06-07-89184-а, рук. И.А. Цыганкова.

Статья поступила в редакцию 16.06.2009.

Труды СПИИРАН. 2009. Вып. 9. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2009. Issue 9. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru