

РАЗРАБОТКА КОМПЬЮТЕРНОЙ СИСТЕМЫ МЕДИЦИНСКОЙ ДИАГНОСТИКИ НА ОСНОВЕ ГЕНЕТИЧЕСКИХ ДАННЫХ И ФАКТОРОВ РИСКА

И. С. Бык

ВВЕДЕНИЕ

Распространенность первичной артериальной гипертензии (АГ) в мире достаточно велика. Важное медико-социальное значение этого заболевания обусловлено высокой распространенностью в общей популяции, и большим количеством осложнений, в том числе со смертельным исходом [1]. Большинство исследований посвящено изучению влияния известных полиморфизмов отдельных генов-кандидатов либо влиянию лишь факторов образа жизни. Таким образом, изучение генетически факторов, характерных для определенных этнических групп при АГ, в совокупности с факторами образа жизни остается актуальным и недостаточно разработанным научным направлением.

Для построения модели классификации пациентов, необходимой для их диагностики [2], необходимо решить две задачи:

- нахождение генетических факторов, ассоциированных с АГ, выступающих в дальнейшем как;
- построение классификационной модели с включением факторов окружающей среды (биологических, поведенческих и психосоциальных) и генетических факторов из предыдущего шага.

МАТЕМАТИЧЕСКИЕ МЕТОДЫ И АЛГОРИТМЫ, ИСПОЛЬЗУЕМЫЕ ДЛЯ ОЦЕНКИ ПРЕДРАСПОЛОЖЕННОСТИ К АГ

Для формирования пространства генетических факторов использовался пакет APSampler [3], который реализует метод Монте-Карло на основе цепей Маркова (Markov Chain Monte-Carlo – MCMC method) с алгоритмом Метрополиса – Гастингса [4].

Формирования генетических факторов. Для генерации выборки $T_1, \dots, T_N$ из вероятностного распределения, заданного плотностью $p(T)$, необходимо, чтобы была задана некоторая марковская цепь с априорным распределением $p_0(T)$ и вероятностями перехода $q_n(T_{n+1} | T_n)$ в момент времени n , а генерация выборки происходит следующим образом:

$$T_1 \sim p_0(T), T_2 \sim q_1(T_2 | T_1), \dots, T_N \sim q_{N-1}(T_N | T_{N-1}) \quad (1)$$

Марковская цепь должна быть однородной и эргодичной, а распределение $p(T)$ инвариантно относительно марковской цепи [2]. Тогда, вне зависимости от начального распределения $p_0(T)$, начиная с некоторого момента времени n выборка, генерируемая по схеме (1), будет выборкой из распределения $p(T)$. Рассмотрим шаг генерации по схеме Метрополиса–Гастингса. Пусть на шаге n сгенерирована конфигурация T_n . На шаге $n+1$ сначала генерируется конфигурация T_* из некоторого предположного распределения $r(T_n | T_*)$. Затем вычисляется величина

$$A(T_*, T_n) = \min \left(1, \frac{p(T_*)r(T_n | T_*)}{p(T_n)r(T_* | T_n)} \right) \quad (2)$$

и точка T_* принимается в качестве следующей точки T_{n+1} с вероятностью $A(T_*, T_n)$. В противном случае, $T_{n+1} = T_n$.

Алгоритмы классификации пациентов на основе логит-модели бинарного выбора. Для классификации пациентов на два класса могут применяться различные скоринговые алгоритмы [5, 6]. Согласно [7, 8], если зависимая переменная Y_i дихотомическая, то, не ограничивая общности, положим, что она принимает бинарные значения: 0, если i -й исход не успешен (признак отсутствует); 1 – иначе. Обозначим вероятность успешного исхода для данных через $\pi(x_i)$. Обычно предполагается, что имеется латентная (не наблюдаемая) переменная Y_i^* , в зависимости от значений которой наблюдаемая переменная

$$y_i = \begin{cases} 1, Y_i^* > 0 \\ 0, Y_i^* \leq 0 \end{cases} \quad (3)$$

Рассмотрим набор p независимых переменных-факторов, обозначаемое вектором $x' = (x_1, x_2, \dots, x_p)$. Скрытая переменная зависит от факторов x в смысле обычной линейной регрессии

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + u \equiv \beta' x + u, \quad x' = (x_1, x_2, \dots, x_p), \beta' = (\beta_0, \beta_1, \dots, \beta_p),$$

где случайная ошибка u имеет распределение $F(z)$, $z \in \mathbb{R}^1$. Тогда

$$\pi(x_i) = P\{Y_i^* > 0 | x_i\} = P\{\beta' x_i + u_i > 0\} = 1 - F(-\beta' x_i) \quad (4)$$

Обозначим $Pr(y=1|x) = \pi(x)$. Множественная модель бинарной логистической регрессии задается уравнением

$$g(x) = \ln \left(\frac{\pi(x)}{1-\pi(x)} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p, \pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (5)$$

РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ

В результате анализа 9 генов-кандидатов, соответствующих АГ, с помощью APSampler с последующим экспертным анализом получены генные факторы, положительно ассоциированные с заболеванием

Таблица 1

Значимые генетические факторы, положительно ассоциированные с АГ

Сочетания для группы больных	OR(95% CI)	p
Т аллель AGT-M235T + AA генотип AGTR2-C3123A	1,57 (1,003..2,45)	0.0292
ТТ генотип AGT -M235T + D аллель ACE-ID	4,14 (1,449..11,836)	0.0019
D аллель eNOS-E298D + C аллель BKR2-T58C	0,416 (0,234..0,739)	0,0018
D аллель eNOS-E298D + D аллель ACE-ID	0,428 (,242..0,755)	0,0022

С учетом этих генных факторов и экзогенных факторов образа жизни, полученных из опросов пациентов РНПЦ “Кардиология” с помощью пакета EViews было построено несколько классификационных моделей. При описании результатов моделирования используются следующие обозначения: AGE – возраст; АО – наличие абдоминального ожирения; BMI – наличие превышенного индекса массы тела; GENDER – пол; MALE_HEREDITY – наличие семейного анамнеза у родственников мужского пола моложе 55 лет; PHIZ – физическая активность; AGT_AGTR2 – генетический фактор.

Модель с генетическими факторами описывается следующим соотношением:

0.468785620686*AGT_AGTR2+1.0923149315*AGE+0.694709982849*АО+1.43011213749*BMI+0.339633016542*PHIZ+1.11919064282*MALE_HEREDITY+0.705012495652*GENDER-1.51265108741

Модель, построенная без учета генетических факторов:

0.997127289299*AGE + 0.74295459272*АО + 0.358463564066*PHIZ+1.43392604062*BMI+0.676926059126*GENDER+1.00733294275*MALE_HEREDITY - 1.39639817108

Все оценки коэффициентов обеих моделей значимы на уровне 0,065. Модели в целом адекватны, но модель, включающая генные факторы обладает лучшей прогностической способностью и качеством: процент

правильных решений 74% и 71% для порога отсечения 0,5; McFadden R-squared 1.93 и 1.83 соответственно.

КОМПЬЮТЕРНАЯ СИСТЕМА ДИАГНОСТИКИ

В результате предыдущих исследований была создана компьютерная система. Она позволяет опрашивать пациентов, сохранять и накапливать данные их визитов и классифицировать их, используя любые применимые к конкретному пациенту модели. КС, с одной стороны, помогает собирать данные для дальнейшего анализа новых комбинаций генов, а с другой дает возможность продемонстрировать пациенту риск развития АГ в его текущем состоянии и как этот риск может измениться, если скорректировать некоторые факторы образа жизни (рисунок 1).

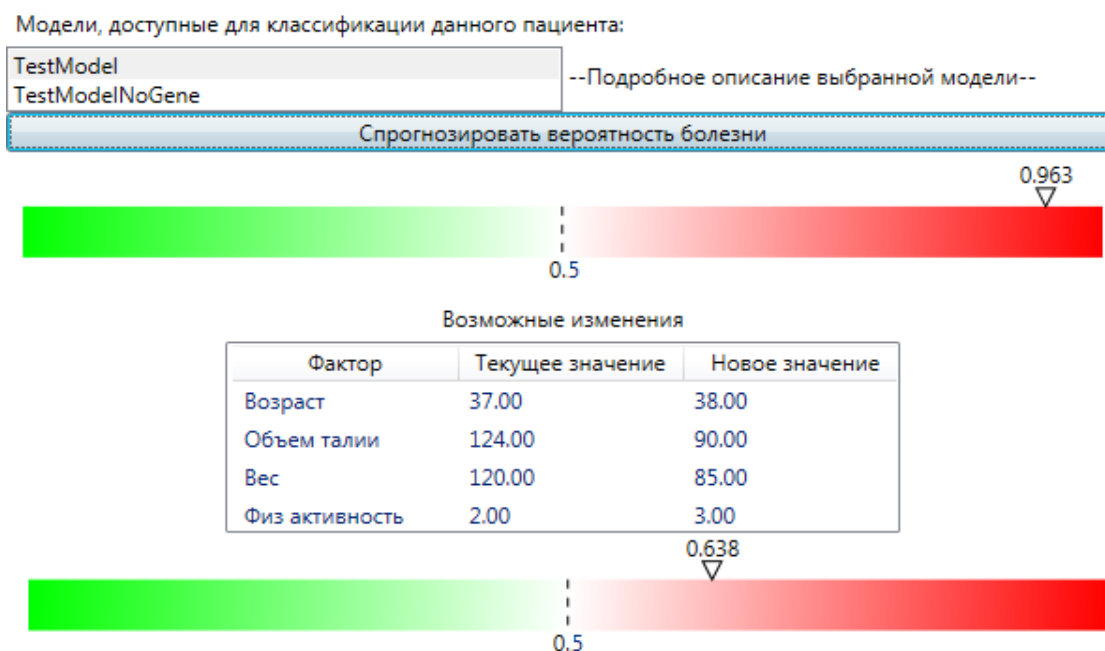


Рисунок 1. Текущие результаты оценки риска заболевания АГ для выбранного пациента, первый – из доступных классификационных моделей и возможный результат после 1 года терапии

ЗАКЛЮЧЕНИЕ

В результате выполнения данной работы: сформировано пространства классификационных признаков (генетических факторов и экзогенных факторов риска); разработан алгоритм классификации на основе логит-модели бинарного выбора; создана программная реализация математической классификационной модели в виде компьютерной системы с

использованием ЯП C# совместно с технологией WPF и сервера баз данных MS SQL Server; проведена апробация КС в РПНЦ “Кардиология”.

Литература

1. *World Health Organization*. Global status report on noncommunicable diseases 2010. Geneva, World Health Organization, 2011.
2. *Pavlova O.S., Malugin V.I., Ogurtsova S.N., Novopolcev A. Yu.* Computer modeling of gene-gene and gene-environment interaction in essential hypertension // ISBRA. Minsk, 2016.
3. *Favorov A.V., Andreewski T.V., Sudomoina M.A., Favorova O.O., Parmigiani G., Ochs M.F.* A Markov Chain Monte Carlo technique for identification of combinations of allelic variants underlying complex diseases in Humans. *Genetics*. –171(4). – 2005. – P. 2113–2121.
4. *Gilks W.R., Richardson S., Spiegelhalter D.I.* Markov Chain Monte Carlo in Practice. London: Chapman & Hall, 1996.
5. *Гринь Н.В., Малюгин В.И.* Исследование точности методов классификации многомерных данных в задачах кредитного скоринга // Вестник ГрГУ. Сер. 2. – 2008. – № 1. – С. 77–85.
6. *Малюгин В.И., Корчагин О.И., Гринь Н.В.* Исследование эффективности алгоритмов классификации заемщиков банков на основе балансовых коэффициентов // Банковский Вестник. – 2009. – № 7. – С. 27–32.
7. *Hosmer D.W.* Applies Logistic Regression. Third Edition / D.W. Hosmer // John Wiley & Sons. – 2013. – 509 p.
8. *Малюгин В.И., Пытляк Е.В.* Оценка устойчивости коммерческих банков на основе эконометрических моделей с дискретными зависимыми переменными // Банковский Вестник. – 2007. – № 4(369). – С. 30–36.

АДАПТИВНЫЕ АЛГОРИТМЫ ПОСТРОЕНИЯ ОПТИМАЛЬНЫХ ПОВЕРХНОСТЕЙ

Д. Д. Васильков

ВВЕДЕНИЕ

Классическим подходом к моделированию поверхностей является двухэтапный метод: построение триангуляции на исходном наборе точек и построение линейных или нелинейных частей поверхности на полученных треугольниках [1]. Однако даже оптимальные алгоритмы построения триангуляции не гарантируют отсутствия треугольников с углами, близкими к 0 или 180 градусам (левая триангуляция на рис. 1). В таком случае качество поверхности существенно ограничивает ее применение на практике.