СТАТИСТИЧЕСКАЯ ОЦЕНКА РИСКА ПЕРВИЧНОЙ АРТЕРИАЛЬНОЙ ГИПЕРТЕНЗИИ НА ОСНОВЕ АЛГОРИТМОВ СТАТИСТИЧЕСКОЙ КЛАССИФИКАЦИИ

Е. Н. Макаревич

Белорусский государственный университет, г. Минск;
eniamak@gmail.com;
науч. рук. – В. И. Малюгин, канд. физ.-мат. наук, доц.

В статье представляются результаты исследования возможности применения алгоритмов статистического и машинного обучения в задачах оценки риска возникновения первичной артериальной гипертензии (ПАГ). На основе логит-модели бинарного выбора исследуется эффективность статистической методики оценки риска ПАГ в режиме продолженного наблюдения за пациентами.

Ключевые слова: риск артериальной гипертензии; статистическое оценивание; алгоритмы статистической классификации; логистическая регрессия.

ВВЕДЕНИЕ

Эффективное решение проблем повышения продолжительности и качества жизни населения на основе современных технологий медицинской диагностики является одним из необходимых условий решения демографической проблемы в целом. В настоящее время набирает популярность превентивная медицина. Такая медицина направлена на предупреждение болезни, а также отличается персональным подходом к каждому больному. Выявление особенностей генотипа позволяет оценить риски развития многих заболеваний, которые в будущем могут реализоваться при определенных условиях, связанных факторами образа жизни и окружающей среды. Регулярный индивидуальный мониторинг за изменениями физиологических показателей позволяет увидеть начало формирования болезни, вовремя взяться за лечение и исключить факторы риска, зависящие от человека. Как следует из статистики Всемирной организацией здравоохранения, в настоящее время основной причиной смертности от болезней во всем мире являются кардиологические заболевания [3]. Одним из наиболее распространенных среди них является первичная артериальная гипертензия. В связи с этим ранняя оценка риска данного заболевания является одной из наиболее актуальных задач превентивной медицинской диагностики.

В [1, 2] был предложен модельный и программный инструментарий для оценки риска первичной артериальной гипертензии на основе генетических факторов и так называемых факторов «окружающей среды», включая биомедицинские и поведенческие факторы. На основе выборки

пациентов, проходивших обследование в Республиканском научно-практическом центре «Кардиология» были построены правила принятия решений относительно состояния пациента («здоров» или «болен»), а также получены статистические оценки риска заболевания в виде апостериорных вероятностей принадлежности к классу больных. Базовой статистической моделью для классификации пациентов была бинарная модель логистической регрессии (логит-модель). Целями данного исследования являются: сравнительный анализ альтернативных алгоритмов дискриминантного анализа, использующих выборку пациентов, прошедших первоначальное обследование; исследование возможности применения базовой логит-модели в режиме продолженного наблюдения за пациентами (follow-up study) на основе выборки пациентов, прошедших повторное обследование через 5 лет; ранжирование факторов риска по степени значимости.

РЕЗУЛЬТАТЫ СРАВНИТЕЛЬНОГО АНАЛИЗА АЛГОРИТМОВ КЛАССИФИКАЦИИ

Начальная выборка, используемая для сравнительного анализа алгоритмов классификации, включает 507 пациентов. Соответствующая ей выборка наблюдений является классифицированной. Предполагается, что истинный номер класса пациента (0 – «здоров» или 1 – «болен») определяется по результатам медицинского обследования. Для целей исследования начальная выборка наблюдений разбита на обучающую (66.8%) и экзаменационную (33.2%). Повторное обследование через 5 лет прошли 204 пациента. В данной выборке 144 пациента принадлежат классу больных и 60 пациентов принадлежат классу здоровых (при первом обследовании: 120 больных и 84 здоровых). Изначально здоровых пациентов, получивших диагноз «болен» при повторном исследовании, – 24 (14 мужчин, 10 женщин). В качестве классификационных признаков используются 2 генетических фактора и факторов «окружающей среды»: возраст, индекс массы тела, показатель абдоминального ожирения, физическая активность, курение, употребление алкоголя. Бинарная целевая переменная указывает на высокий риск (значение целевой переменной равно 1) либо на незначительный риск (значение целевой переменной равно 0) развития данной болезни. Для сравнительного анализа используются следующие алгоритмы классификации: деревья решений и их ансамбли; линейный дискриминантный анализ, алгоритм *K*-ближайших соседей, байесовский классификатор, модель логистической регрессии бинарного выбора (логит-модель). Результаты экспериментов представлены в табл. 1.

Таблица 1

Сравнительный анализ эффективности алгоритмов классификационных

Факторы	Алгоритм классификации	Точность	Чувствительность
Без генетических факторов	Дерево решений	0.77	0.86
Все факторы	Логит-модель	0.80	0.86
Все факторы	Байесовский классификатор	0.75	0.76

В качестве критериев эффективности алгоритмов используются показатели, принятые в медицинской диагностике: «чувствительность», «специфичность» и «точность», основанные на оценках вероятностей ошибок классификации: условных P_0, P_1 – для класса здоровых и больных соответственно и безусловной P. (см. формулы в табл. 2). Согласно табл. 1 среди рассматриваемых алгоритмов предпочтительнее выглядит алгоритм классификации на основе логит-модели. Данный алгоритм используется далее для классификации пациентов прошедших повторное обследование через 5 лет. Результаты классификационных экспериментов для логит-модели (порог отсечения 0.62) представлены в табл. 2 (где π_0 – доля здоровых пациентов, d и d_0 – номера классов на основе агоритма классификации и медицинских обследованиях).

Таблица 2

Условные и безусловная вероятности ошибок классификации и основанные на них метрики

Метрики	Оценки вероятностей	
	Начальная выборка	Повторная выборка
$\pi_0(\pi_1 = 1-\pi_0)$	0.332	0.294
$P_0 = \Pr(d = 1 \mid d_0 = 0)$	0.291	0.75
$P_1 = \Pr(d = 0 \mid d_0 = 1)$	0.249	0.062
Точность = $1-P$, $P = \pi_0 P_0 + (1-\pi_0)P_1$	0.737	0.735
Чувствительность = $1 - P_1$	0.751	0.938
Специфичность = $1 - P_0$	0.709	0.250

Низкое значение метрики специфичности в таблице 2 для повторной выборки, полученной в режиме продолженного наблюдения, может объясняться усилением факторов риска, которым соответствуют классификационные признаки: появился лишний вес и абдоминальное ожирение, уменьшилась физическая активность, у некоторых пациентов появилась

привычка к курению. Такие изменения в состоянии пациентов привели к их отнесению к классу пациентов с более высоким риском заделывания.

С помощью алгоритма классификации на основе деревьев решений был исследован относительный вклад каждого фактора риска в результат классификации. На рис. 1 представлены результаты исследований, из которых следует, что наиболее значимыми факторами, влияющими на результат классификации, являются следующие факторы: абдоминальное ожирение, индекс массы тела (индикатор избыточности веса), возраст пациента.

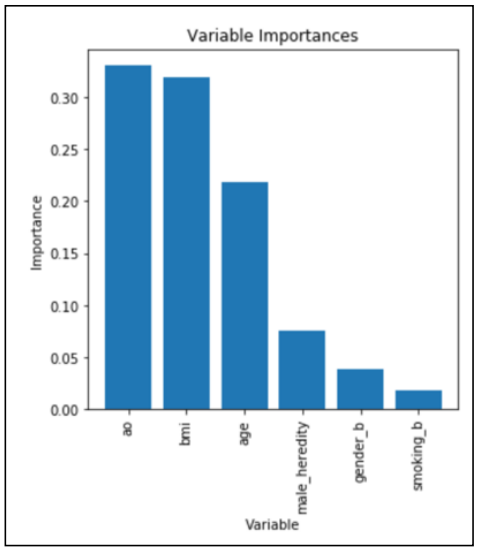

Рис. 1. Относительный вклад признаков в предсказание риска заболевания на различных множествах признаков

Библиографические ссылки

1. *Павлова О. С., Малюгин В. И.* Полигенные ассоциации полиморфизма генов ренин-ангиотензин альдостероновой системы при эссенциальной артериальной гипертензии // Артериальная гипертензия. 2016. № 22 (3). С. 253–261.
2. *Pavlova O. S., Malugin V. I.* Computer Analysis of Essential Hypertension Risk on the Base of Genetic and Environmental Factors // Proc. of the 11th Intern. Conf. «Computer Data Analysis and Modeling», Minsk, 2016. P. 289–293.
3. The top 10 causes of death [Electronic resource] // World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (date of access: 24.05.2018).
4. An Introduction to Statistical Learning: with Applications in R [Electronic resource] / G. James [et al.]. URL: https://www.academia.edu/36691506/An_Introduction_to_Statistical_Learning_Springer_Texts_in_Statistics_An_Introduction_to_Statistical_Learning (date of access: 24.05.2018).
5. *Kuhn M., Johnson K.* Applied Predictive Modeling. M., 2013.