УДК 519.2

© 2023 г. **М.З. Ермолицкая,** канд. биол. наук
(Институт автоматики и процессов управления ДВО РАН, Владивосток;
Владивостокский государственный университет)

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ЗАБОЛЕВАЕМОСТИ РАКОМ ЖЕЛУДКА

С помощью методов машинного обучения разработаны прогностические модели заболеваемости раком желудка на территории Приморского края по данным за период с 2007 г. по 2021 г.: полиномиальная регрессия, ридж-регрессия и искусственные нейронные сети. Наиболее эффективна, согласно оценкам качества, модель, обученная с помощью ИНС.
Ключевые слова: методы машинного обучения, главные компоненты, регрессионный анализ, искусственные нейронные сети, рак желудка.

DOI: 10.22250/18142400_2023_77_3_41

Введение

Методы математической статистики получили широкое применение в биомедицине благодаря росту и доступности компьютерных технологий. Первичный анализ данных с расчетом описательных статистик, оценки достоверности различий в выборках, обнаружением взаимосвязей между показателями, выявлением закономерностей активно используется при различных исследованиях. При этом выбор статистических критериев зависит от типов имеющихся данных и от поставленных задач. Чаще всего используются непараметрические методы анализа данных, основанные на том, что количественные показатели не подчиняются нормальному закону распределения [1, 2].

Разработка и модификация прогностических моделей имеет важное значение в современном здравоохранении. Применение методов машинного обучения и искусственного интеллекта позволяет прогнозировать заболеваемость населения с учетом факторов риска и территориальной принадлежности, определять вероятность развития того или иного исхода, рассчитывать

отношение шансов и т.п. [3-8] Поиск подходящей модели, как правило, сводится к построению нескольких отдельных моделей, выбору из них наиболее эффективной либо созданию на их основе комбинаторной модели.

К традиционным методам прогнозирования можно отнести регрессионный анализ (полиномиальная, экспоненциальная, логистическая, ридж-регрессия и др.), используемый для установления взаимосвязи между зависимой и независимыми переменными, построения уравнения зависимости и оценки его качества [9 – 13]. Для выявления скрытых линейных и нелинейных связей в больших и сложных наборах данных применяют нейросетевые методы [14 – 17].

Целью данного исследования является разработка прогностической модели заболеваемости раком желудка с применением методов машинного обучения.

Методы и результаты

В качестве отклика прогностической модели были взяты стандартизованные показатели заболеваемости раком желудка (РЖ) на территории Приморского края за период с 2007 г. по 2021 г. В качестве предикторов использовали показатели качества жизни населения, сгруппированные в три главные компоненты согласно предыдущему исследованию [18]. В первую компоненту (PC_1) вошли социально-экономические показатели, во вторую (PC_2) – социально-гигиенические показатели (гигиена, загрязнение окружающей среды), в третью (PC_3) – показатели потребления продуктов питания. Выделенные главные компоненты объясняют 72% общей вариации исходных данных. Коэффициенты корреляции между заболеваемостью РЖ и главными компонентами показывают значимую сильную связь: 0.86, 0.77, 0.75 соответственно.

Построить качественную модель с учетом всех трех главных компонент с помощью методов регрессионного анализа с линейной функцией не удалось, не все коэффициенты регрессии значимы.

В данном исследовании для построения прогностической модели использовали три метода машинного обучения: полиномиальную регрессию, ридж-регрессию и искусственные нейронные сети (ИНС). Построение моделей осуществляли в программе RStudio (Version 1.0.136). Все показатели предварительно были стандартизованы.

Полиномиальная регрессия позволяет учесть нелинейность отклика:

$$\hat{y} = \beta_0 + \beta_1(PC_3 * PC_2 * PC_1) + \beta_2(PC_3 * PC_2 * PC_1)^2,$$

где $\hat{y}$ – предсказанные значения; β – коэффициенты регрессии.

Для аппроксимации данных использовали ортогональный полином Чебышева второй степени (функция poly{stats}). В полученном уравнении свободный коэффициент регрессии равен 9.387e-17. Значение мало, им можно пренебречь. Остальные коэффициенты значимы согласно критерию Стьюдента: β_1 = -2.780 (p-value = 2.44e-05), β_2 = 2.039 (p-value = 0.0004). При этом доля дисперсии зависимой переменной, объясняемая моделью, составляет 84.9%. Остатки нормально распределены согласно критерию Шапиро – Уилка (p-value = 0.879).

Ридж-регрессия (гребневая регрессия) реализует классическую регуляризацию Тихонова, при которой оценки параметров модели β находят из условия минимизации [19]:

$$\beta = \arg\min[\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m}\beta_j x_{ij})^2 + \lambda\sum_{i=1}^{m}\beta_j^2],$$

где y – наблюдаемые значения отклика; x_{ij} – предикторы (в нашем случае главные компоненты (PC_1, PC_2, PC_3)); λ – параметр сглаживания.

В матричном виде гребневая регрессия представляет собой линейную модель:

$$\hat{y} = H_y X,\ H = X(X^T X + \lambda I)^{-1} X^T,$$

где λI – диагональная матрица, называемая "гребнем".

При $\lambda \to 0$ регуляризованное решение стремится к МНК-решению, т.е. обычной линейной модели. Подбор λ осуществляли с помощью обобщенной перекрестной проверки на основе критерия GCV (generalized cross-validation):

$$GCV = \frac{1}{n}\sum_i \left(\frac{y_i - \hat{y}_i}{1 - tr(H)/n}\right)^2.$$

Выбор модели проводился путем поиска минимального значения критерия GCV в серии из 20 моделей со значениями от 0 до 2, с шагом 0.1 (функция lm.ridge{MASS}). Полученное значение λ, которое минимизирует критерий GCV, равно 0.023. Коэффициенты регрессии окончательной модели с оптимальным значением λ представлены в табл. 1.

Таблица 1

Полученные значения коэффициентов ридж-регрессии

	PC_1	PC_2	PC_3
Коэффициенты регрессии	0.886	0.142	0.082

Доля дисперсии зависимой переменной, объясняемая ридж-моделью, составляет 74%. Остатки нормально распределены (p-value=0.511).

Для обучения искусственной нейронной сети, позволяющей воспроиз-

водить сложные зависимости, использовали модель математического нейрона Мак-Каллока – Питтса:

$$S = \sum_{i=1}^{n} w_i x_i,$$

где S – линейная комбинация входных сигналов (адаптивный сумматор); x_i- входные сигналы (предикторы); w_i - синаптические весы, выражающие важность соответствующих входных сигналов для выходного сигнала.

ИНС можно представить как систему, состоящую из ряда сильно взаимосвязанных узлов, называемых «нейронами», которые организованы в слои, обрабатывающие информацию с использованием динамических откликов состояния на внешние входные данные (входные сигналы).

Обучение ИНС с использованием обратного распространения ошибки осуществляли с помощью функции neuralnet{neuralnet}. Функция позволяет выполнять гибкие настройки посредством индивидуального выбора ошибки и функции активации. В качестве функции активизации использовали линейную регрессию (linear.output=TRUE), в качестве функции ошибок рассчитывали среднеквадратическую ошибку (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Настройка ИНС осуществлялась экспериментально. Рассматривали многоуровневые структуры с числом нейронов на каждом слое от 1 до 6. Коэффициенты матрицы весов на первом шаге обучения сети инициализировались случайным образом. Поиск оптимальной сети осуществлялся в цикле с изменением случайного числа (seed.current). Обучение сводилось к оптимальному подбору коэффициентов матрицы весов для минимизации функции ошибок.

В результате получили наилучшую модель, состоящую из четырех слоев с разным количеством нейронов в каждом слое (hidden=c(6,2,4,6)) при случайном числе – 3 (рис. 1).

Проверка адекватности обученной ИНС заключалась в расчете коэффициента детерминации и среднеквадратической ошибки (табл. 2).

Таблица 2

	Коэффициент детерминации (R^2)	Среднеквадратическая ошибка (MSE)
Полиномиальная регрессия	0.85	0.14
Ридж-регрессия	0.74	0.24
ИНС	0.97	0.025

Остатки нейросетевой модели подчиняются нормальному закону распределения (p-value = 0.223).

Сравнивая полученные модели по оценкам качества, можно выделить модель ИНС, предсказанные значения которой ближе к наблюдаемым значениям (рис.2).

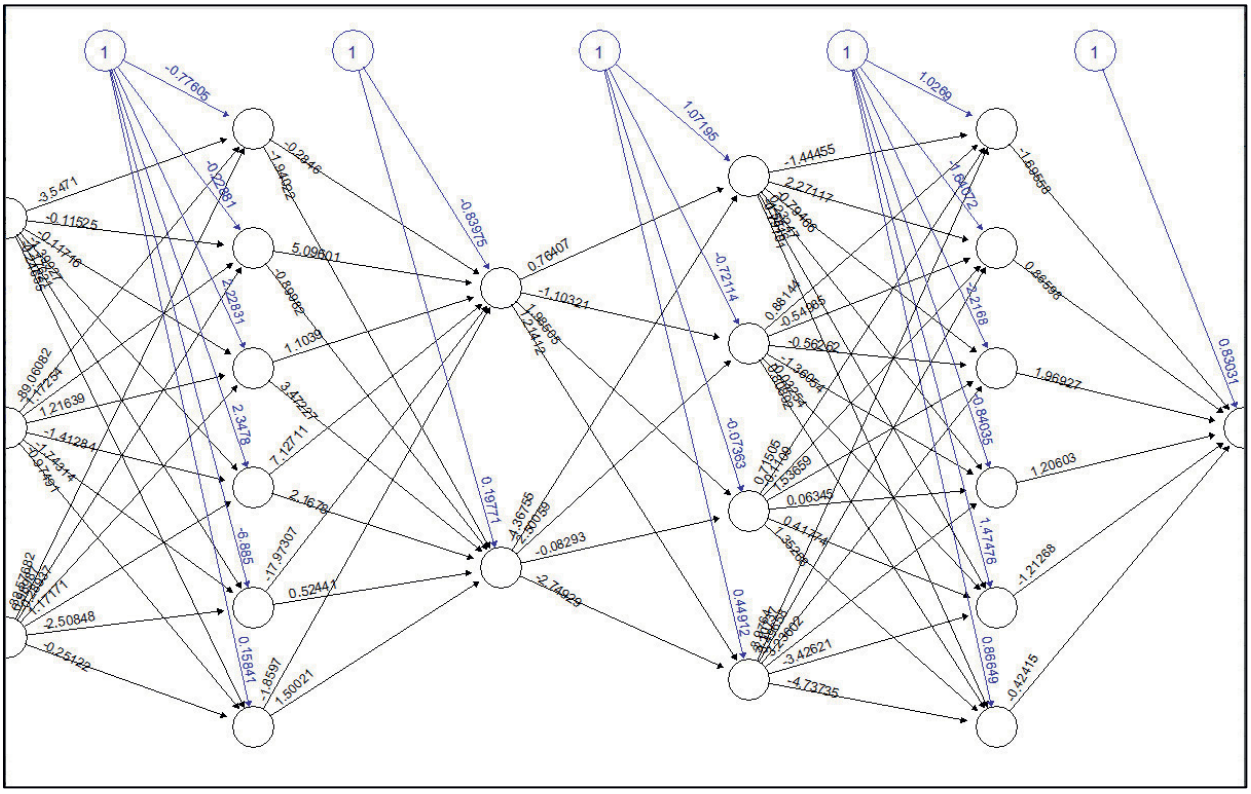

Рис. 1. График обученной искусственной нейронной сети.

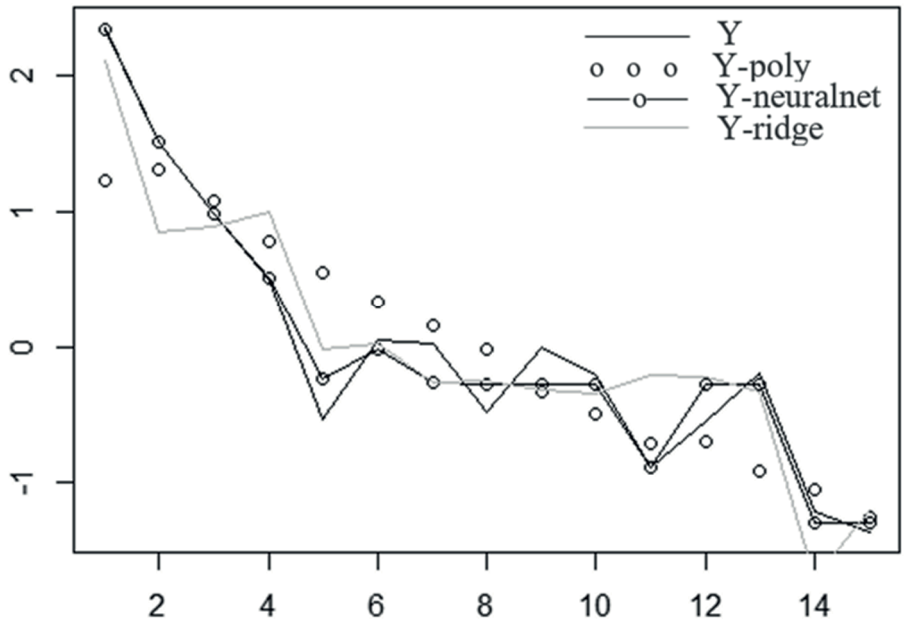

Рис. 2. Наблюдаемые и предсказанные значения полиномиальной регрессии (Y-poly), ридж-регрессии (Y-ridge) и ИНС (Y-neuralnet).

Заключение

В результате данного исследования получены три модели, позволяю-

щие прогнозировать заболеваемость раком желудка на территории Приморского края. Наилучшая модель, согласно оценкам качества, построена с помощью нейросетевого метода. Обучение ИНС проходило на всей выборке данных. Деление на обучающую и тестовую выборки не проводилось в связи с малым количеством наблюдений. Дальнейшее исследование подразумевает сбор данных для увеличения выборки, проведение кросс-валидации и проверку полученных на данном этапе результатов.

Применение нейросетевого метода дает возможность эффективно решать задачи прогнозирования динамики заболевания раком желудка по показателям качества жизни населения Приморского края.

ЛИТЕРАТУРА

1. *Румянцев П.О., Саенко У.В., Румянцева У.В.* Статистические методы анализа в клинической практике. – Ч. I. Одномерный статистический анализ // Проблемы эндокринологии. – 2009. – Т. 55, № 5. – С. 48–55.
2. *Survival* of stomach cancer patients in n Western Kazakhstan: a registry-based study / Tulyayeva A.B., Bekmuhamedov Y.J., Zhamalieva L.M., et al. // Human Ecology. – 2021. – № 1. – P. 51–56.
3. *Загоруйченко А.А., Карпова О.Б.* Актуальные подходы к прогнозированию и моделированию заболеваемости населения в России (обзор) // Санитарный врач. – 2022. – № 8. – С. 596–606.
4. *Кондратьев М.А.* Методы прогнозирования и модели распространения заболеваний // Компьютерные исследования и моделирование. – 2013. – Т. 5, № 5. – С. 863–882.
5. *Лучинин А.С.* Прогностические модели в медицине // Клиническая онкогематология. Фундаментальные исследования и клиническая практика. – 2023. – Т. 16, № 1. – С. 27–36.
6. *Ляпин В.А., Маренко В.А., Елохова Ю.А., Ложников А.Е.* Построение моделей на основе эмпирических данных по онкозаболеваемости населения // Информатика и системы управления. – 2023. – № 1(75). – С. 28–36.
7. *Van Smeden M., Reitsma J.B., Riley R.D., Collins G.S., Moons K.Gm.* Clinical prediction models: diagnosis versus prognosis // J Clin Epidemiol. – 2021. – № 132. – P. 14–25.
8. *Басова Л.А., Карякина О.Е., Кочорова Л.В., Мартынова Н.А.* Роль прогностических моделей в повышении качества медицинских услуг // Фундаментальные исследования. – 2013. – № 9(3). – С. 323 – 326.
9. *Ляпин В.А., Маренко В.А.* Модели функциональных зависимостей демографических показателей и экологических характеристик // Информатика и системы управления. – 2020. – № 3(73). – С. 3–12.
10. *Климушкин А.В., Борщук Е.Л., Бегун Д.Н., Бегун Т.В., Куланова А.М.* Прогноз заболеваемости злокачественными новообразованиями в Оренбургской области // Современные проблемы науки и образования. – 2021. – № 2. – С. 1 – 16.
11. *Юдин С.В.* Эпидемиологический анализ онкологической заболеваемости как показатель здоровья населения Приморского края //Тихоокеанский медицинский журнал. – 2006. – № 3. – С. 43–45.

12. *Ермолицкая М.З.* Прогнозирование заболеваемости раком молочной железы с применением обобщенной аддитивной модели // Информатика и системы управления. – 2022. – № 3(73). – С. 25–30.
13. *Sekeroglu B., Tuncal K.* Prediction of cancer incidence rates for the European continent using machine learning models // Health Informatics Journal. – 2021. – 27(1):1460458220983878.
14. *Sidey-Gibbons Jenni A. M., Sidey-Gibbons Chris J.* Machine learning in medicine: a practical introduction // BMC Medical Research Methodology. – 2019. – Vol. 19(1). – P. 1–18.
15. *De Hond A.A.H., Leeuwenberg A.M., Hooft L., Kant I.M.J., et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review // NPJ Digit Med. – 2022. – № 5(1). – P. 1–13.
16. *Волчек Ю.А., Шишко О.Н., Спиридонова О.С., Мохорт Т.В.* Положение модели искусственной нейронной сети в медицинских экспертных системах // Медицинские науки. – 2017. – № 9. – С. 4–9.
17. *Хасанов А.Г., Шайбаков Д.Г., Жернаков С.В. и др.* Нейронные сети для прогнозирования динамики развития заболеваний // Креативная хирургия и онкология. – 2020. – № 10(3). – С. 198–204.
18. *Ермолицкая М.З., Кику П.Ф.* Выявление взаимосвязи между показателями качества жизни населения и заболеваемостью злокачественными новообразованиями в Приморском крае // Здоровье населения и среда обитания – ЗНиСО. – 2022. – Т. 30, № 6. – С. 7–14.
19. *Мастицкий С.Э., Шитиков В.К.* Статистический анализ и визуализация данных с помощью R. – М.: ДМК Пресс, 2015.

Статья представлена к публикации членом редколлегии А.И. Абакумовым.

E-mail:
Ермолицкая Марина Захаровна – ermmz@mail.ru.