УДК 519.2

© 2020 г. **М.З. Ермолицкая**, канд. биол. наук
(Институт автоматики и процессов управления ДВО РАН, Владивосток)

ИСПОЛЬЗОВАНИЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ПРОГНОЗИРОВАНИЯ ПОЛОЖИТЕЛЬНОЙ ДИНАМИКИ ПРИ ЛЕЧЕНИИ БОЛЬНЫХ ТУБЕРКУЛЕЗОМ

С использованием программы RStudio разработана нейросетевая модель, прогнозирующая положительную динамику при лечении больных в стационаре туберкулезного диспансера. Точность (accuracy) представленной модели на тестовой выборке составляет 99.4%, значение среднеквадратической ошибки – 0.013.

Ключевые слова: статистический анализ данных, искусственная нейронная сеть, больные туберкулезом.

DOI: 10.22250/isu.2020.66.139-147

Введение

Применение современных методов статистического моделирования и машинного обучения дает возможность разрабатывать и модифицировать прогностические модели, способные существенно помочь при решении сложных задач в здравоохранении и медицине [1]. Существуют разные подходы к прогнозированию инфекционных заболеваний, их применение зависит от количества и качества исходных данных [2 – 5]. Поэтому предварительно следует изучить имеющиеся данные и определить степень их влияния на исследуемый процесс. Для этого чаще всего используют корреляционный анализ. Поиск подходящей модели, как правило, сводится к построению нескольких отдельных моделей, выбору из них оптимальной либо созданию на их основе комбинаторной модели. Отдельно следует выделить применение искусственных нейронных сетей (ИНС), которые могут быть использованы в качестве диагностического инструмента для прогнозирования заболевания и поддержки в расширении роли компьютерных технологий в диагностике для оперативного лечения [6, 7]. Основное преимущество ИНС является их спо-

собность извлекать скрытые линейные и нелинейные связи даже в больших и сложных наборах данных.

Для прогнозирования распространенности туберкулеза и моделирования лечения больных туберкулезом эффективны такие традиционные методы как регрессионный анализ (полиномиальная, экспоненциальная модели [8], логистическая модель [9] и др. [10, 11]) и современные методы интеллектуального анализа – искусственные нейронные сети [12 – 14]. Сравнение моделей, полученных с помощью этих методов, позволяет с высокой степенью точности выделить наиболее качественную модель, пригодную для практического применения.

В данной работе представлена разработанная с помощью искусственных нейронных сетей модель, позволяющая предсказать наличие положительной динамики процесса выздоровления больных туберкулезом при стационарном лечении в диспансере. Проведено сравнение ранее полученной регрессионной модели с нейросетевой моделью на основе рассчитанных оценок качества и среднеквадратической ошибки.

Статистическая обработка и анализ данных

Исходная выборка данных по лечению больных туберкулезом в Приморском краевом противотуберкулезном диспансере (ГЗУБ «ПКТД») состояла из 507 наблюдений и 78 показателей, характеризующих образ жизни людей (вредные привычки), диагноз, сопутствующие заболевания, дополнительное обследование, медикаментозное лечение, приобретенные заболевания, динамику лечения. В ходе разведочного анализа категориальные данные были кодированы. Показатели с большим количеством отсутствующих наблюдений исключены из рассмотрения. Для выявления значимых показателей, существенно влияющих на положительную динамику выздоровления пациентов, были использованы следующие критерии: критерий Шапиро – Уилка, согласно которому распределение данных не является нормальным; критерий Манна – Уитни для выявления различий по категории пол (процесс выздоровления протекает одинаково у мужчин и женщин) и метод Тау Кендалла для определения зависимостей между показателями. В результате было выделено 20 показателей (495 наблюдений), которые в разной степени влияют на процесс выздоровления больных туберкулезом. Значимые коэффициенты корреляции показателя «положительная динамика» с другими показателями представлены в табл. 1. Эти показатели использовали для построения прогностических моделей. Результаты первичной обработки и анализа исходных данных представлены в работе [15].

Таблица 1

Количество койко-дней	Вес пациента	Работающий	ИБС	МБТ	Изониазид	Этамбутол	Стрептомицин	Амикацин	Протионамид	Циклосерин	Изменение режима	Поражение ЖКТ	Гепатоксичность	Негативация мокроты	Исчезновение лабораторных признаков	Регрессия рентгенологических проявлений	Лечение завершено (стационарный этап)	Лечение прервано	Амбулаторное лечение (интенсивная фаза)
0.436	0.109	0.184	–0.127	–0.097	0.138	0.152	0.113	0.098	–0.108	–0.095	–0.258	0.096	0.124	0.478	0.921	0.849	0.509	–0.475	–0.507

Построение нейросетевой модели

Искусственные нейронные сети – это мощный метод моделирования, позволяющий воспроизводить чрезвычайно сложные зависимости. Пользователь нейронной сети отбирает и подготавливает данные, выбирает нужную архитектуру сети, а затем запускает алгоритм обучения, который автоматически воспринимает структуру данных. Решение в сети формируется множеством простых нейроноподобных элементов, образующих граф со взвешенными синаптическими связями, которые совместно и целенаправленно работают на получение общего результата. Главным строительным блоком такой сети, согласно модели математического нейрона Мак-Каллока – Питтса, является искусственный нейрон, основная функция которого – сформировать выходной сигнал y в зависимости от входных сигналов $x_{1...}x_n$. Значения входных сигналов могут усиливаться или ослабляться в зависимости от знака синаптических весов $w_{1,...},w_n$ [16]

$$\mathrm{s} = \sum_{i=1}^{n} w_i x_i \ , \qquad (1)$$

где S – линейная комбинация входных сигналов (адаптивный сумматор).

Выходной сигнал сумматора поступает в нелинейный преобразователь F с функцией активизации. Здесь функция активизации имеет логистический вид (сигмоид), так как зависимая переменная, отражающая эффективность лечения больных туберкулезом, является бинарной переменной:

$$\mathrm{F(s)} = \frac{e^s}{1+e^s} \ . \qquad (2)$$

После преобразования результат подается на выход (рис. 1).

В общем случае искусственная нейронная сеть состоит из трех основ-

ных компонент: входного слоя, скрытого (вычислительного) слоя и выходного слоя. Для всех искусственных нейронных сетей присущ принцип параллельной обработки сигналов, выполняемый путем объединения большого числа нейронов различной конфигурации в скрытом слое и их последующей обработке.

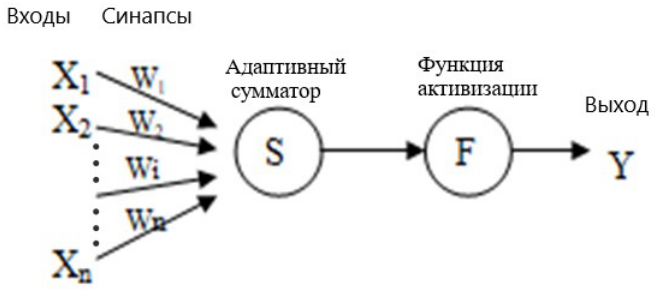

Рис. 1. Структура искусственного нейрона.

При построении сети для решения задач классификации взвешенные комбинации Y выходного слоя представляют собой прогноз, который указывает на принадлежность распознаваемого объекта к определенной группе. Если распознаются только два класса, то в выходном слое персептрона находится только один элемент, который обладает двумя реакциями – положительной и отрицательной, т.е. на выходе значение сигмоидального нейрона лежит в интервале [0,1].

Для обучения искусственной нейронной сети в программе RStudio использовали функцию neuralnet() из пакета neuralnet, позволяющую создавать множество внутренних слоев в сети. В качестве входных сигналов (предикторов модели) рассматривали 20 показателей, выделенных на основе корреляционного анализа (табл. 1). Предварительно исследуемая выборка была поделена на обучающую и тестовую в стандартном соотношении: 3/4 наблюдений для обучающей выборки и 1/4 – для тестовой.

Настройка искусственной нейронной сети осуществлялась экспериментально. Рассматривали одно- и двухуровневые структуры, с числом нейронов на каждом слое от 2 до 12. Коэффициенты матрицы весов на первом шаге обучения сети инициализировались случайным образом. Поиск оптимальной сети осуществлялся в цикле с изменением случайного числа (seed.current) в диапазоне от 1 до 50000. Обучение сводилось к оптимальному подбору коэффициентов матрицы весов для минимизации функции ошибок (функции потерь). Функция ошибок используется для расчета ошибки между реальными и полученными данными. Основная цель – минимизировать эту ошибку.

В качестве функции ошибок рассчитывали среднеквадратическую

ошибку (MSE и RMSE):

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

$$RMSE = \sqrt{MSE}\,, \qquad (3)$$

где y_i – наблюдаемые значения; $\hat{y}_i$ – предсказанные значения.

Чем ближе значение среднеквадратической ошибки к нулю, тем лучше построенная модель.

В итоге получили наилучшую нейросетевую модель с минимальной среднеквадратической ошибкой (MSE = 0.013, RMSE = 0.1140175) на тестовой выборке, состоящую из пяти нейронов в одном слое при случайном числе seed.min = 13 (рис. 2).

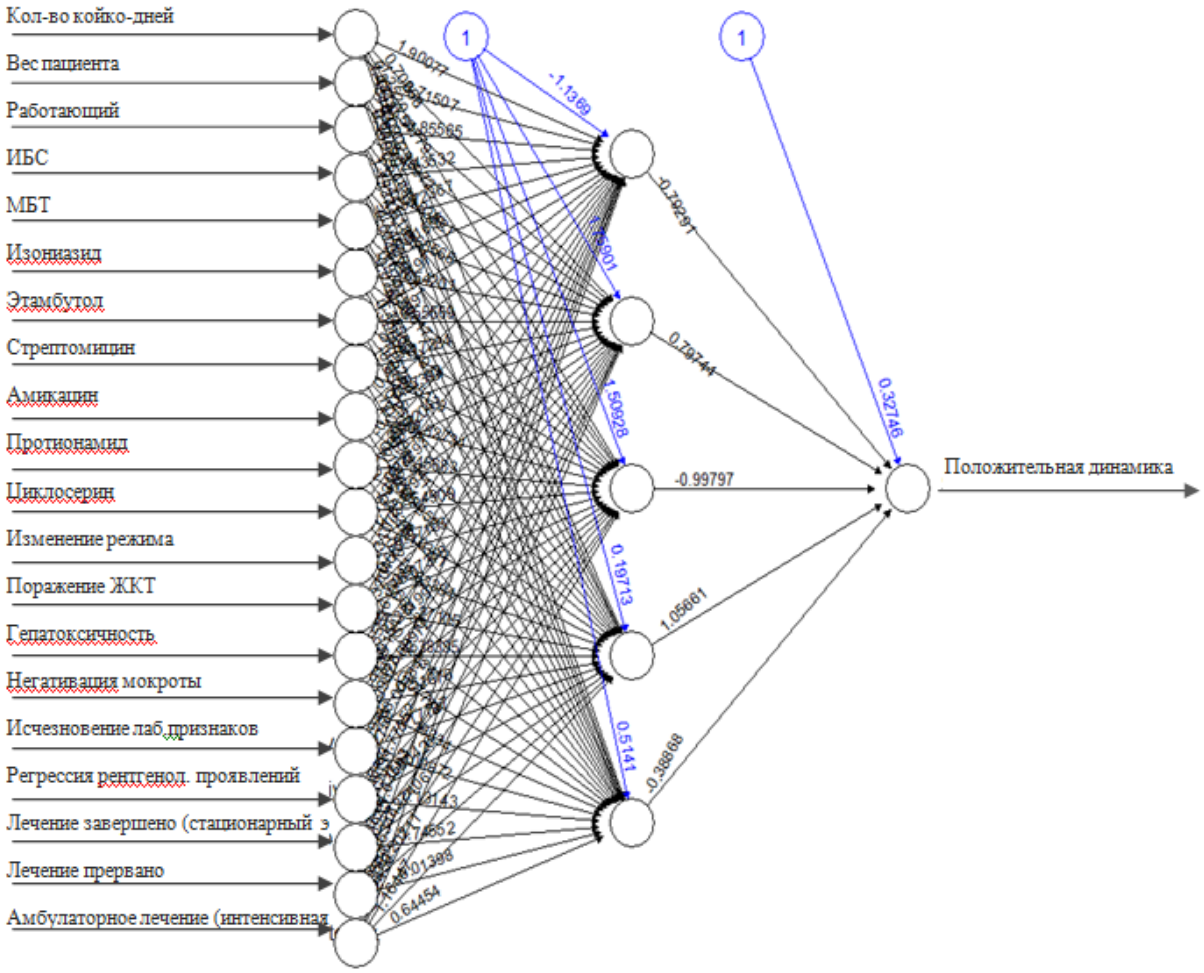

Рис.2. График обученной искусственной нейронной сети.

На рис. 2 видно, что на выходной слой сети, состоящий из одного элемента, обладающего двумя реакциями (положительной и отрицательной), поступает 5 сигналов, представляющих собой комбинацию значений исходных переменных. Эта взвешенная комбинация и есть прогноз, указывающий на принадлежность распознаваемого объекта к определенной

группе – с наличием положительной динамики лечения или ее отсутствием.

В программе RStudio при использовании функции neuralnet() прогноз строится с помощью функции compute().

Для построения матрицы неточностей (Confusion Matrix) на тестовой выборке использовали функцию table(y,ŷ). Результат представлен в табл. 2.

Таблица 2

Наблюдаемые значения	Прогноз	
	0	1
0	25	0
1	1	114

Один объект тестовой выборки предсказан неверно.

Точность модели (accuracy) рассчитывалась как доля правильно классифицированных объектов (количество правильно классифицированных объектов к общему числу объектов):

$$accuracy = \frac{\sum_{i=1}^{n} I[\hat{y}_i = y_i]}{n}. \quad (4)$$

Полученное значение точности модели (accuracy) на тестовой выборке высоко и составляет 99,4%. При этом чувствительность модели (Sensitivity) равна проценту верно предсказанных позитивных исходов

(114/(114 + 1))*100% = 99.1%.

Специфичность модели (Specificity) показывает процент верно предсказанных негативных исходов

(25/(25 + 0))*100% = 100%.

Построенная нейросетевая модель адекватно описывает процесс выздоровления больных туберкулезом. Высокое качество модели говорит о правильности выбора структуры сети и значимых предикторов.

В результате анализа исходных данных выделения двадцати значимых показателей были построены две прогностические модели – логистическая регрессионная и нейросетевая. В процессе создания логит-модели предикторы с коэффициентами регрессии, уровень значимости для которых по критерию Стьюдента больше 0.05 (p-value > 0.05), постепенно были удалены из рассмотрения. Конечная модель включает девять значимых предикторов, характеризующих дополнительное обследование, медикаментозное лечение и приобретенные заболевания.

Построенная нейросетевая модель состоит из пяти нейронов в одном слое и содержит двадцать предикторов на входе. Для определения качества моделей на тестовой выборке были рассчитаны следующие оценки: средне-

квадратические ошибки и точность предсказания (табл. 3).

Таблица 3

Тип модели	Среднеквадратическая ошибка (MSE)	Квадратный корень из среднеквадратической ошибки (RMSE)	Точность (accuracy), %	Чувствительность (Sensitivity), %	Специфичность (Specificity), %
Логистическая регрессионная	0.01428	0.11952	98.57	98.26	100
Нейросетевая	0.013	0.1140175	99,4	99.1	100

Полученные оценки свидетельствуют о высоком качестве разработанных моделей и при сравнении позволяют выделить наилучшую, адекватную модель с наименьшей ошибкой и наибольшим значением точности, т.е. построенную с помощью искусственных нейронных сетей. При этом количество предикторов в логистической регрессионной модели значительно меньше, что бывает очень важным при выборе вида модели для построения прогноза в медицине.

Заключение

Применение двух классических методов машинного обучения для решения задачи прогнозирования положительной динамики при лечении больных туберкулезом дает возможность с помощью разных подходов к моделированию значений бинарного отклика разрабатывать модели, оценивать их и, сравнивая полученные результаты, выбирать наилучшую. При использовании методов множественной логистической регрессии и искусственных нейронных сетей для определения качества построенных моделей рассчитывались одинаковые оценки: среднеквадратическая ошибка (MSE) и точность предсказания (accuracy, Sensitivity, Specificity). Полученные значения этих оценок позволяют на основе их сравнения выделить наиболее эффективную модель.

«Увы, универсального способа прогнозирования заболеваемости не существует – оптимальный подход следует выбирать, сравнивая результаты, полученные с помощью различных техник на основе эмпирических данных. Зачастую сложно отдать предпочтение одному методу прогнозирования – несколько подходов дают результаты сопоставимого качества», – отмечал М.А. Кондратьев [2]. Поэтому подобные исследования подразумевают использование различных методов прогнозирования, с предварительным выбором значимых показателей и последующим сравнением полученных мо-

делей посредством построения матрицы неточностей и расчета ошибок прогнозирования.

В нашем случае наилучший прогноз получен с помощью нейросетевой модели. Поскольку ошибки прогнозирования ничтожно малы и лишь один объект тестовой выборки предсказан неверно, можно считать полученную модель эффективной и пригодной для использования в качестве вспомогательного инструмента для оперативного лечения больных туберкулезом в условиях стационара.

ЛИТЕРАТУРА

1. *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.
2. *Кондратьев М.А.* Методы прогнозирования и модели распространения заболеваний // Компьютерные исследования и моделирование. – 2013. – Т. 5, № 5. – С. 863-882.
3. *Azeez A., Obaromi D., Odeyemi A., Ndege J., Muntabayi R.M.* Seasonality and Trend Forecasting of Tuberculosis Prevalence Data in Eastern Cape, South Africa, Using a Hybrid Model // Environmental Research and Public Health. – 2016. – № 13(8).
4. *Brookspollock E., Cohen T., Murray M.* The impact of realistic age structure in simple models of tuberculosis transmission // PLoS ONE. – 2010. – Vol. 5(1). URL: https://doi.org/10.1371/journal.pone.0008479 (accessed 12.05.2020)
5. *Menzies N.A., Wolf E., Connors D.* Progression from latent infection to active disease in dynamic tuberculosis transmission models: A systematic review of the validity of modelling assumptions // Lancet Infect. Dis. – 2018. – Vol. 18(8). – P. 226-236.
6. *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.
7. *Smith L.* An Introduction to Neural Networks // Unpublished draft, University of Stirling, 2001. URL: http://www.cs.stir.ac.uk/~lss/NNIntro/InvSlides.html (accessed 12.05.2020)
8. *Wang J., Wang C., Zhang W.* Data analysis and forecasting of tuberculosis prevalence rates for smart healthcare based on a novel combination model // Applied sciences. – 2018. – 8(9). URL: https://doi.org/10.3390/app8091693 (accessed 12.05.2020)
9. *Mello F.C.Q., Bastos L.G.V., Soares S.L.M., Rezende V.M., Conde M.B., Chaisson R.E.* Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: a cross-sectional study // BMC Public Health. – 2006. – Vol. 6(43). URL: https://doi.org/10.1186/1471-2458-6-43 (accessed 12.05ю2020)
10. *Мун С.А., Глушов А.Н., Штернис Т.А., Ларин С.А., Максимов С.А.* Регрессионный анализ в медико-биологических исследованиях. – Кемерово: КемГМА, 2012.
11. *Aguiar F.S., Almeida L.L., Ruffino-Netto A., Kritski A.L., Mello F.C., Werneck G.L.* Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients // BMC Pulm. Med, 2012. URL: http://doi: 10.1186/1471-2466-12-40 (accessed 12.05.2020)
12. *Dande P., Samant P.* Acquaintance to artificial neural networks and use of artificial intelligence as a diagnostic tool for tuberculosis: a review // Tuberculosis. – 2018. – Vol.108. –

P.1-9.
13. *Fojnica A., Osmanovice A., Badnjevice A.* Dynamical model of tuberculosis-multiple strain prediction based on artificial neural network // 5th Mediterranean Conference on Embedded Computing (MECO), (Piscataway, NJ: IEEE). – 2016. – P. 290-293.
14. *Khan M.T., Kaushik A.Ch., Ji L., Malik S.I., Ali S., Wei D.* Artificial neural networks for prediction of tuberculosis disease // Front. Microbiol, 2019. URL: https://doi.org/10.3389/fmicb.2019.00395 (accessed 12.05.2020)
15. Ермолицкая М.З. Перспективы вылечиться от туберкулеза. Анализ данных средствами программы RStudio // Информатика и системы управления. – 2020. – 1(63). – С. 50-58.
16. *Шитиков В.К., Мастицкий С.Э.* Классификация, регрессия и другие алгоритмы DataMining с использованием R, 2017. URL: https://github.com/ranalytics/data-mining. (дата обращения 12.05.2020)

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

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