УДК 51-76

© 2024 г. **А.Б. Кригер**[1], канд. физ.-мат. наук,
А.А. Яковлев[1,2], канд. физ.-мат. наук
[1] (Дальневосточный федеральный университет, Владивосток)
[2](Институт автоматики и процессов управления ДВО РАН, Владивосток)

ВЫДЕЛЕНИЕ ИНФОРМАТИВНЫХ ПРЕДИКТОРОВ ИЗ ЦИТОКИНОВОГО ПРОФИЛЯ ПАЦИЕНТА

По данным концентрациям цитокинов проведено исследование связей между предикторами. Разработан алгоритм выделения наиболее информативных цитокинов. По данному алгоритму выделены информативные предикторы из цитокинового профиля пациентов. Проведена оценка качества классификации и устойчивости результатов.

Ключевые слова: вирус папилломы человека, рак шейки матки, цитокиновый профиль, машинное обучение, логистическая регрессия, моделирование.

DOI: 10.22250/18142400_2024_80_2_32

Введение

В статье [1] рассмотрены вопросы применения нейронной сети для прогнозирования диагноза пациента на основе цитокинового профиля. Цитокины участвуют во всех этапах иммунного ответа и являются индикаторами состояния иммунной системы индивида [2, 3]. В настоящее время определение цитокинов достаточно активно используется в клинической практике, так как их количественные характеристики и соотношения между собой отражают динамику патологического процесса, коррелируют с активностью заболевания и позволяют судить об эффективности проводимого лечения [3].

Моделирование показало хорошие результаты, однако очень остро встает финансовый вопрос. Стоимость анализа биоматериала на один цитокин в лабораториях г. Владивостока показана в [5, 6]. Она составляет, в зави-

Работа выполнена в рамках государственного задания FWFW-2022-0002.

симости от лаборатории, порядка 900 руб. + стоимость забора биоматериала. Так как в моделировании участвует 8 цитокинов, то не каждая женщина финансово способна сделать такой анализ, что, в свою очередь, влияет на женское здоровье и репродуктивность. В связи с этим возник вопрос: можно ли уменьшить количество исследуемых цитокинов без потери точности прогноза или с минимальными потерями? В нашем исследовании мы попытались ответить на этот вопрос.

Проверка гипотез различия между группами пациентов. Отбор предикторов для классификации пациентов

Предмодельный статистический анализ данных приведен в [1], не будем его повторять. Отметим лишь основные тезисы:

1. Данные не подчиняются закону нормального распределения.

2. Анализ с применением критерия Манна–Уитни показал, что между контрольной и онкологической группами в некоторых цитокинах есть различия.

3. Корреляционный анализ не выявил сильных связей между цитокинами в выделенных группах пациентов.

Тестирование гипотез относительно существенных различий между пациентами с диагнозом и без подтвержденного диагноза (по группам) не выявило предикторов, которые бы однозначно классифицировали пациентов всех пяти представленных групп. Более того, результаты тестов показали наличие противоречий.

Результаты оценивания критерия Манна–Уитни указывали на существенное различие пациентов с диагнозом и пациентов контрольных (1 и 2 групп) по некоторым цитокинам. Ситуация с группой 3 неоднозначна, поэтому были сделаны следующие шаги:

во-первых, принято решение разделить пациентов на группы: «здоровые», с диагнозом и «риском»;

во-вторых, для выявления наиболее информативных цитокинов (далее предикторов) авторы применили модель логистической регрессии (пошаговый алгоритм).

Такой подход позволяет получить не только достоверные оценки статистической значимости предикторов, но и провести содержательный анализ их влияния на диагноз.

В результате предпринятых действий удалось получить статистически значимые модели классификации, с высоким качеством разделения и понижением размерности задачи.

Для отбора предикторов был применен следующий алгоритм.

Агрегирование пациентов в укрупненные группы: «условно здоровый», «условно больной».

Оценка логистических моделей – со всеми предикторами: модель, полученная методом пошаговой регрессии, – только со статистически значимыми предикторами. Поскольку все переменные нормализованы, модели в том числе позволяют оценить степень влияния каждого из предикторов на результат.

Оценка качества классификации и устойчивости результата для полученных моделей. Отбор предикторов.

Проверка результатов методом кросс-валидации. Было проведено агрегирование пациентов в укрупненные группы: «условно здоровый», состоящую из контрольной, первой и второй групп; и «условно больной», состоящую из третьей и четвертой групп. Результаты оценивания моделей и обоснование отбора предикторов представлены в табл 1.

Таблица 1

Показатель (переменная)	Коэффициент	Стандартная ошибка	*p-value*
Модель со всеми доступными переменными			
(Intercept)	1.2199	0.3519	0.0005
IL-1	-1.5953	0.6213	0.0102
TNF-α	-0.4154	0.2935	0.1569
IL-6	0.1555	0.5425	0.7743
IL-10	0.7401	0.6124	0.2269
IL-18	-0.5043	0.3436	0.1422
IL-8	-0.5677	0.3612	0.1161
VEGF	-1.1153	0.5655	0.0486
TGF-β1	1.5953	0.6157	0.0096
возраст	-0.2490	0.4193	0.5526
Итоговая модель, уточненная пошаговым исключением			
(Intercept)	1.1469	0.3052	0.0001
IL-1	-1.5607	0.4784	0.0011
TNF-α	-0.5219	0.2593	0.0442
IL-8	-0.7806	0.3177	0.0140
TGF-β1	1.0137	0.4082	0.0130

Для итоговой модели все переменные статистически значимы на уровне 0.01 (TNF-α значимость 0.04, что для медицинских исследований, как правило, слишком слабый уровень), при сохранении существенной прогностической точности.

На рис. 1, 2 представлены ROC-кривые, соответствующие моделям,

которые позволяют оценить качество прогноза на основе полученных моделей. Как видно из рисунков, прогностические качества исходной модели (со всеми переменными) и модели пониженной размерности различаются слабо.

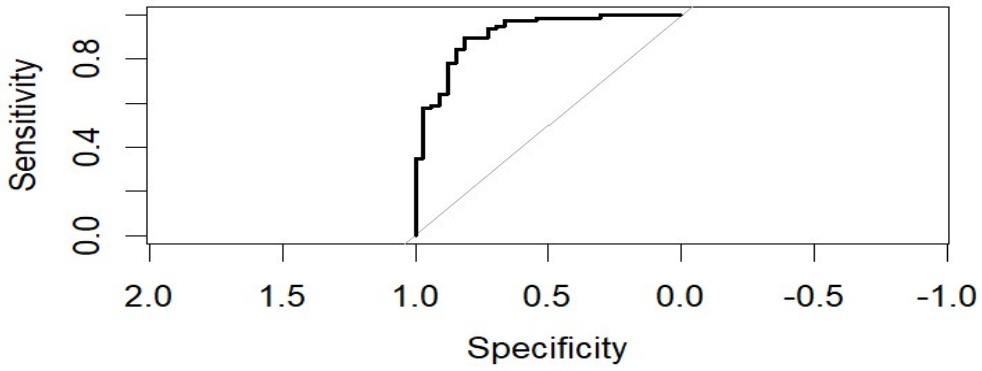

Рис. 1. ROC-кривая для модели со всеми переменными (auc = 0.91).

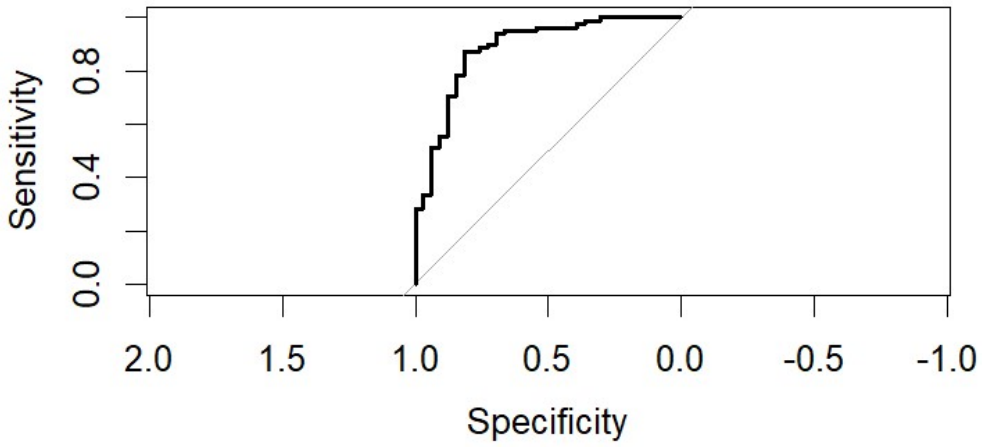

Рис. 2. ROC-кривая для модели после пошагового исключения (auc = 0.88).

Адекватность модели пониженной размерности подтверждается в том числе следующим:

сохраняются знаки коэффициентов модели;

значения для отобранных предикторов показывают, что они вносят наибольший вклад в предсказание вероятности.

Вопрос о возникновении смещенности в оценке коэффициентов не актуален, так как в нашем случае задача анализа степени вклада предикторов не стоит.

Оценка результатов

Для понижения размерности задачи и классификации пациентов мы использовали логистическую регрессию, которая является важным методом в области искусственного интеллекта и машинного обучения [7, 8]. Оценка

качества обученной модели и устойчивости прогнозирования реализована методом кросс-валидации с разбиением данных на обучающую и тестовую в разных пропорциях [9]. Вычислительная процедура для каждого варианта разбиения была сделана 50 раз и рассчитано среднее значение. Результаты вычислений были сведены в табл. 2 (AUC обозначает площадь под ROC-кривой).

Таблица 2

Вариант разбиения train / test	Чувствительность	Специфичность	Gmean*	Точность	AUC
80% / 20%	0.85	0.78	0.81	0.82	0.87
75% / 25%	0.85	0.78	0.81	0.79	0.88
70% / 30%	0.84	0.76	0.79	0.81	0.87

Результаты классификации пациентов по выделенным цитокинам (табл. 2) показывают высокую устойчивость при различных способах разбиения. Средние значения AUC варьируют в пределах 0.01. ROC-кривые при различных способах разбиения показаны на рис. 3-5. Поскольку в таблице приведены усредненные по результатам кросс-валидации показатели качества классификации, на рисунках представлены ROC-кривые для отдельных итераций, приблизительно соответствующие усредненным показателям.

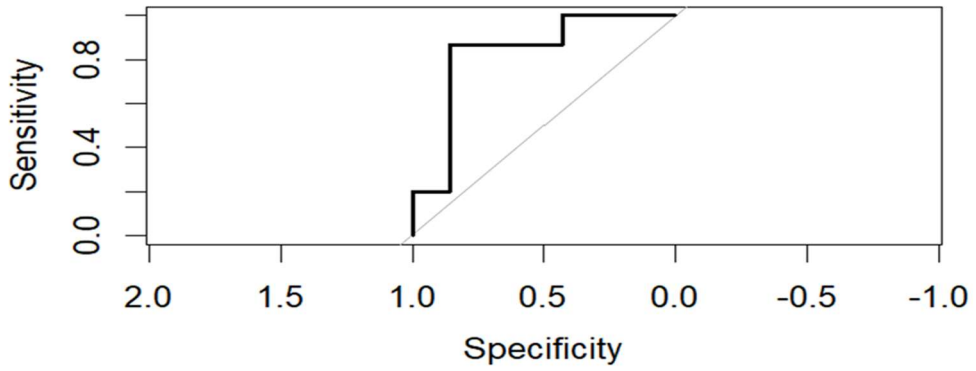

Рис. 3. ROC-кривая для разбиения train / test 80% / 20% (auc = 0.83).

Представленные выше результаты (как количественные, так и инфографика) позволяют утверждать, что классификация пациентов (разбиение на две группы – с подтвержденным диагнозом, с высоким риском и «условно» здоровых) достоверно определяется всего четырьмя цитокинами. Причем результат не зависит от соотношения обучающей и тестовой выборок.

* Gmean=$\sqrt{sensetivity \cdot specifity}$ – среднее геометрическое показателей чувствительности и специфичности, стандартный сбалансированный показатель для оценки эффективности классификации [10].

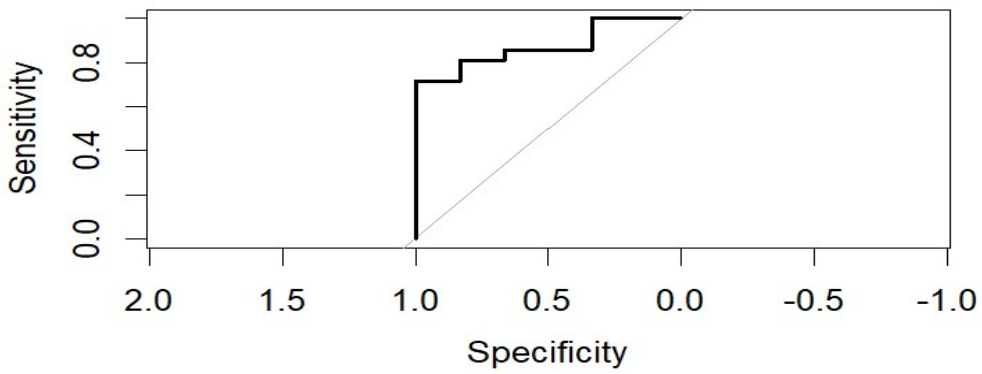

Рис. 4. ROC-кривая для разбиения train / test 75% / 25% (auc = 0.87).

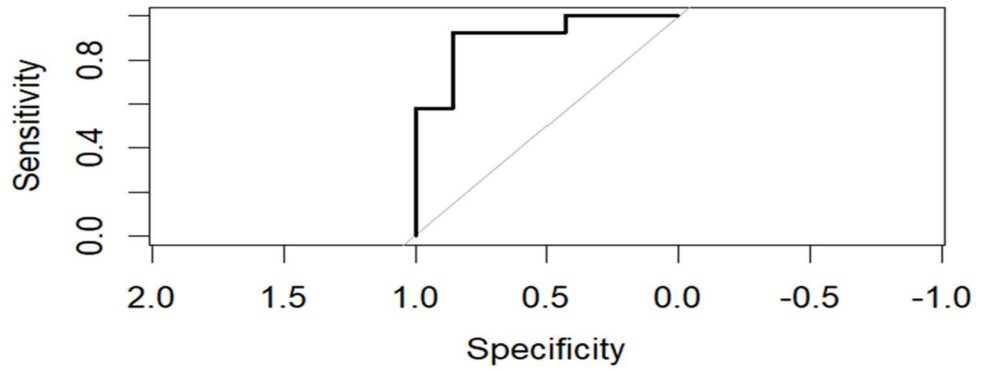

Рис. 5. ROC-кривая для разбиения train / test 70% / 20% (auc=0,89).

Мы попробовали изменить агрегированные группы пациентов, переведя третью группу пациентов в условно здоровую. После этого было проведено моделирование по методике, описанной выше. Результаты были сведены в табл. 3.

Таблица 3

Вариант разбиения train / test	Чувствительность	Специфичность	Gmean	Точность	AUC
80% / 20%	0.684	0.97	0.79	0.92	0.96
75% / 25%	0.69	0.97	0.80	0.92	0.95
70% / 30%	0.65	0.97	0.78	0.90	0.95

Несмотря на то, что при подобном разбиении площадь под ROC-кривой больше, можно заметить, что чувствительность существенно падает. Падение чувствительности связано с выбором одного и того же «порогового» уровня классификации для обоих методов агрегации (использовался очень жесткий порог в 70% вероятности с целью получения «убедительной» классификации пациентов). Причина выбора одинаковых «пороговых» уровней классификации именно в получении сопоставимых результатов. Однако оптимальные пороговые значения, различные для моделей, выбира-

ются именно исходя из соотношения «Чувствительности» и «Специфичности» (они должны быть сбалансированы), что мы и наблюдаем в нашем случае.

Необходимо отметить, что даже несмотря на ухудшение параметров «Чувствительность» и «Специфичность / Gmean», полученные предикторы показывают высокую устойчивость моделирования при различных способах разбиения обучающей и тестовой выборок. Это свидетельствует о правильности определения предикторов.

Для сравнения результатов было проведено моделирование со всеми цитокинами. Результат моделирования сведен в табл. 4.

Таблица 4

Вариант разбиения train / test	Чувствительность	Специфичность	Gmean	Точность	AUC
80% / 20%	0.82	0.78	0.79	0.81	0.86
75% / 25%	0.82	0.78	0.79	0.80	0.87
70% / 30%	0.81	0.76	0.78	0.80	0.86

Результаты, приведенные в табл. 4, подтверждают, что уменьшение числа предикторов в модели практически не изменяют результат классификации. Показатели «качества» классификации ожидаемо чуть более сбалансированы.

Заключение

В ходе исследования из восьми цитокинов (IL-1, TNF-α, IL-6, IL-10, IL-18, IL-8, VEGF, TGF-β1) цитокинового профиля пациента нами были отобраны четыре наиболее значимых: IL-1, TNF-α, IL-8, TGF-β1. Моделирование показало, что агрегирование пациентов в укрупненные группы «условно здоровый», состоящую из контрольной, первой и второй групп, и «условно больной», состоящую из третьей и четвертой групп, дает результаты лучше, чем если бы третья группа пациентов была в составе группы «условно здоровые». При этом и в первом, и во втором случае мы наблюдаем устойчивость показателей при различных методах разбиения агрегированных групп на обучающую и тестовую выборки. Это показывает, что предиктивные цитокины были определены верно.

Полученные результаты моделирования позволяют снизить количество исследуемых цитокинов для анализа биоматериала на вирус папилломы человека, который может перерасти в рак шейки матки. Это способствует уменьшению стоимости анализов и увеличению охвата женщин для обследования, что в свою очередь поможет определить наличие вируса папилло-

мы в организме и возможную онкологию на ранней стадии развития, чтобы как можно скорее приступить к лечению.

Авторы благодарят д.м.н., профессора Е.В. Маркелову и аспирантку ТГМУ М.А. Черникову за предоставленные данные.

ЛИТЕРАТУРА

1. *Кузьмин З.Д., Яковлев А.А.* Применение нейронной сети для ранней диагностики рака шейки матки на основе цитокинового профиля пациента // Информатика и системы управления. – 2024. – № 1 (79). – С. 35–45.
2. *Цитокины* как индикаторы состояния организма при инфекционных заболеваниях. Анализ экспериментальных данных / А.А. Яковлев, А.И. Абакумов, А.В. Костюшко, Е.В. Маркелова // Компьютерные исследования и моделирование. – 2020. – Т. 12, № 6. – С. 1409–1426.
3. *Роль* цитокин-опосредованных механизмов в развитии посттравматического остеомиелита нижней челюсти // А.Б. Кригер, Е.В. Паскова, Е.В. Маркелова и др. – Медицинская иммунология. – 2019. – Т. 21, № 5. – С. 953–958.
4. *Цитокиновый* статус доноров крови и ее компонентов / Г.А. Зайцева, О.А. Вершинина, О.И. Матрохина и др. // Фундаментальные исследования. – 2011. – № 3. – С. 61–65.
5. https://asklepiy-dv.ru/services/laboratory/immunologicheskie-issledovaniya/ (дата обращения 29.02.2024 г.).
6. https://unilab.su/services/analyses/vladivostok/156/57344/ (дата обращения 29.02.2024 г.).
7. https://aws.amazon.com/ru/what-is/logistic-regression/ (дата обращения 29.03.2024 г.).
8. https://loginom.ru/blog/logistic-regression-roc-auc (дата обращения 29.03.2024 г.).
9. https://wiki.loginom.ru/articles/cross-validation.html (дата обращения 03.04.2024 г.).
10. https://en.wikipedia.org/wiki/Geometric_mean (дата обращения 19.04.2024 г.)

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

E-mail:
Кригер Александра Борисовна – kriger.ab@dvfu.ru;
Яковлев Анатолий Александрович – yakovlev-aa@iacp.dvo.ru.