ПОИСК ЭФФЕКТИВНОГО АЛГОРИТМА КЛАСТЕРИЗАЦИИ ДЛЯ ДИАГНОСТИКИ ПРОТОКОВОЙ АДЕНОКАРЦИНОМЫ ПОДЖЕЛУДОЧНОЙ ЖЕЛЕЗЫ НА РАННИХ СТАДИЯХ

Ю. С. Борисова

Саратовский национальный исследовательский государственный университет им. Н. Г. Чернышевского, Россия
E-mail: jboris5873@yandex.ru

Как известно за последнее время в области искусственного интеллекта был сделан технологический прорыв, движущим фактором которого стали нейронные сети. Нейронные сети являются одним из разнообразного количества инструментов машинного обучения. Различают два вида машинное обучение: контролируемое и неконтролируемое, к последнему относят кластеризацию. Кластеризацию используют как предварительный этап для других методов анализа, например, классификация или деревья решений, а также как самостоятельный инструмент анализа данных.

В данной работе для упрощения диагностики рака была применена кластеризация, являющаяся методом интеллектуального анализа данных, основная особенность которых заключается в установлении наличия и характера скрытых закономерностей. Протоковая аденокарцинома поджелудочной железы – одно из самых опасных и смертельных раковых заболеваний этого органа, поэтому ее диагностика на ранних стадиях сможет сократить смертность пациентов. Из этого следует, что чем точнее алгоритм распределяет данные на кластеры, тем лучше проводится прогнозирование биомаркеров как предикторов протоковой аденокарциномы поджелудочной железы.

В работе сравнивались такие алгоритмы кластеризации как иерархическая кластеризация, метод k-средних и DBSCAN с целью обнаружения наиболее точного и эффективного алгоритма. В ходе исследования было установлено, что метод k-средних является наиболее эффективным алгоритмом для диагностики протоковой аденокарциномы на ранних стадиях.

FINDING AN EFFECTIVE CLUSTERING ALGORITHM FOR DETECTING PANCREATIC DUCTAL ADENOCARCINOMA IN THE EARLY STAGES

J. S. Borisova

As we know, there has been a recent technological breakthrough in the field of artificial intelligence, driven by neural networks. Neural networks are one of the machine learning tools variety. There are two types of machine learning: supervised and unsupervised, the latter includes clustering. Clustering is used as a preliminary step for other analysis methods, such as classification or decision trees, or as a stand-alone data analysis tool. In this paper, clustering, which is a data mining technique whose main feature is to establish the presence and nature of hidden patterns, was applied to simplify cancer diagnosis. Pancreatic ductal adenocarcinoma is one of the most dangerous and deadly cancers of this organ, so its early diagnosis will be able to reduce the mortality of patients. It follows that the more accurately the algorithm distributes data into clusters, the better the prediction of biomarkers as predictors of pancreatic ductal adenocarcinoma. The paper compared clustering algorithms such as hierarchical clustering, k-means method and DBSCAN to find out the most accurate and efficient algorithm. The k-means method was found to be the most effective algorithm for early diagnosis of ductal adenocarcinoma.

Благодаря развитию медицины и технологий за последние 60 лет уровень смертности в мире упал в два раза и в 2021 году составлял около 8 смертей на 1 тыс. населения. Несмотря на это онкология занимает лидирующие позиции по количеству смертей в мире. По данным ВОЗ в 2020 году первое место по числу жизней, унесенных различными раковыми заболеваниями, число которых составило более 5,8 млн – занимают страны Азии. Второе место занимает Европа – 1,96 млн. В России от онкологии скончалось 291,5 тыс. человек.

Чаще всего рак диагностируется после 60 лет – самая большая группа заболевших (17,9%) приходится на возраст 65-69. По оценкам ВОЗ, каждый пятый мужчина и каждая шестая женщина на планете заболеют раком на каком-либо этапе жизни. В 2020 году онкологический диагноз был поставлен более чем 19 млн человек.

Протоковая аденокарцинома поджелудочной железы (PDAC) является наиболее распространенным неопластическим заболеванием поджелудочной железы, составляющим более 90% всех злокачественных опухолей этого органа. На сегодняшний день PDAC четвертая по частоте причина смертности от рака во всем мире. Средняя продолжительность жизни больных таким типом рака составляет 5-6 месяцев, и только 11% людей в Соединенных Штатах живут более 5 лет после установки диагноза [1]. В России каждый год число заболевших раком поджелудочной железы практически совпадает с числом умерших от него, что делает эту болезнь по-настоящему роковой.

Ранние симптомы неспецифичны и часто носят периодический характер, поэтому более 80% случаев диагностируются на поздних стадиях, когда опухоль уже распространилась локально или метастазировала в другие органы. Улучшение раннего выявления PDAC значительно повлияло бы на прогноз пациентов, поскольку сообщалось о выживаемости более 60% после случайной диагностики опухолей, когда они все еще ограничивались поджелудочной железой и были менее 2 см [2].

Клинические данные получены из нескольких центров: Банка тканей поджелудочной железы Barts, Университетского колледжа Лондона, Университета Ливерпуля, Испанского национального центра исследований рака, больницы Кембриджского университета и Белградского университета.

Ключевыми характеристиками являются четыре биомаркера анализа мочи: креатинин, LYVE1, REG1B и TFF1. Креатинин – это белок, часто используемый в качестве показателя функции почек. LYVE1 (рецептор 1 гиалуроновой кислоты эндотелия лимфатических сосудов) – белок, который может играть роль в метастазировании опухоли. REG1B – белок, связанный с регенерацией поджелудочной железы. TFF1 – фактор трилистника 1, связанный с регенерацией и восстановлением мочевыводящих путей [3].

Возраст и пол, включенные в набор данных, могут играть роль в выявлении рака поджелудочной железы. Набор данных включал также несколько других биомаркеров, которые, были собраны не у всех пациентов, поэтому в обработку данных не включались. В ходе исследования проводили анализ выбранных методов кластеризации: иерархической кластеризации, метода K-средних и

DBSCAN, с использованием языка Python.

Проведенная кластеризация была отображена на графиках распределения точек данных на кластеры. Анализ кластеров показал, что у определенных кластеров наблюдаемый максимум превышал норму креатинина в моче в 0,27-2,17 мг/мл. Однако повышенное содержание креатинина само по себе может сигнализировать о проблемах с мочевыводящими путями, в которые PDAC часто дает метастазы. Более красноречивым фактором является LYVE1. Его повышенное содержание в моче сигнализирует о распространении опухоли, так как LYVE-1 представляет собой рецептор клеточной поверхности на лимфатических эндотелиальных клетках.

Также некоторые кластеры, в которых предполагалось скопление данных о пациентах с карциномой, имели большой разброс значений REG1B. Семейство регенерирующих (Reg) белков, к которым относится REG1B, представляет собой группу лектиноподобных белков С-типа, обнаруженных у пациентов с панкреатитом и во время регенерации островков поджелудочной железы. Повышение уровня Reg1A и Reg1B наблюдается у пациентов по мере прогрессирования доброкачественного эпителия протоков от интраэпителиальной неоплазии поджелудочной железы (PanIN) до протоковой аденокарциномы. У пациентов с PDAC уровни Reg1A и Reg1B в сыворотке всегда значительно выше, чем у здоровых людей [4].

Такой же вывод был сделан и в отношении количества TFF1. Белки TFF1 выделяются клетками желудочно-кишечного тракта и участвуют в защите от повреждения его слизистой оболочки и ее последующем восстановлении. Стоит учесть, что в поджелудочной железе белки трилистника экспрессируются в основном в раковых или сильно поврежденных клетках.

Обобщая установленные в работе предположения на основе анализа графиков результатов кластеризации данных по раку поджелудочной железы было доказано качество кластеризации при помощи математических расчётов: скорректированного индекса Рэнда, коэффициента Силуэта.

Следует отметить, что индекс Рэнда оценивает, насколько много из тех пар элементов, которые находились в одном кластере, и тех пар элементов, которые находились в разных кластерах, сохранили это состояние после кластеризации. То есть эта метрика показывает, насколько результат кластеризации с помощью одного метода похож на результат кластеризации с помощью другого метода. В свою очередь коэффициент Силуэта показывает, насколько объект похож на свой кластер по сравнению с другими кластерами.

Таким образом, чем ближе значение коэффициента или индекса к 1, тем эффективней была проведена кластеризация.

На основании данных табл. 1 был сделан вывод, что результаты кластеризации с помощью K-средних похожи на результаты иерархической кластеризации больше чем на результаты DBSCAN, так как их индекс Рэнда равен 0,637. Значит DBSCAN точно не подходит на роль эффективного алгоритма для анализа медицинских данных биомаркеров рака поджелудочной железы. Были про-

ведены дальнейшие исследования и рассчитан коэффициент Силуэта, результаты которого показаны в табл. 2.

Таблица 1

Матрица результатов расчета индекса Рэнда

Алгоритм кластеризации	К-средних	Иерархическая	DBSCAN
К-средних	1	0,637	0,327
Иерархическая	0,637	1	0,177
DBSCAN	0,327	0,177	1

Таблица 2

Результаты расчета коэффициента Силуэта

Алгоритм кластеризации	Рак поджелудочной железы
К-средних	0,385
Иерархическая	0,341
DBSCAN	0,305

Из табл. 2 видно, что рассчитанный коэффициент Силуэта имеет наибольшее значение 0,385 у К-среднего, а это подтверждает результаты расчета индекса Рэнда.

Таким образом, для кластеризации данных скрининга протоковой аденокарциномы поджелудочной железы наиболее эффективным является алгоритм кластеризации К-средних.

СПИСОК ЛИТЕРАТУРЫ

1. *Siegel R. L., Miller K. D., Fuchs H. E., Jemal A.* Cancer statistics // CA Cancer Journal for Clinicians. 2022. № 72 (1). С. 7-33.

2. *Orth M., Metzger P., Mayerle J.* Pancreatic ductal adenocarcinoma: biological hallmarks, current status, and future perspectives of combined modality treatment approaches // Radiation Oncology. 2019. № 14 (1). С. 1-20.

3. *Debernardi S., Blyuss O., Rycyk D., Srivastava K., Jeon C. Y.* Urine biomarkers enable pancreatic cancer detection up to 2 years before diagnosis // International Journal of Cancer. 2023. №. 152 (4). С. 769-780.

4. *Debernardi S., O'Brien H., Algahmdi A. S., Crnogorac-Jurcevic T., Malats N., Stewart G. D.* A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case-control study // PLoS Medicine. 2020. № 17 (12).