Научная статья

Классификатор бактериального и вирусного воспаления по анализу гемограмм с использованием методов машинного обучения

Андрей Андреевич Венерин[1]✉, Николай Игоревич Каневский[2]

[12]Институт клинической медицины им. Н.В. Склифосовского, Сеченовский университет, г. Москва, Российская Федерация

[12]Институт цифровой медицины, Сеченовский университет, г. Москва, Российская Федерация

[1]venerin.andrey@gmail.com, http://orcid.org/0000-0002-8960-5772

[2]k_okin@mail.ru, http://orcid.org/0000-0003-4322-0110

Аннотация. В работе представлен вариант упрощенного гематологического классификатора, группирующий объекты по группам нормы, бактериального и вирусного воспаления. В работе представлены реализации методов машинного обучения с учителем: решающее дерево и случайный лес. Полученные результаты могут лечь в основу написания полноценного классификатора-помощника врача, существенно ускоряющего процесс анализа гемограмм пациентов.

Ключевые слова: машинное обучение, гемограммы, воспаление, случайный лес, дерево решений

Для цитирования: Венерин А.А., Каневский Н.И., Классификатор бактериального и вирусного воспаления по анализу гемограмм с использованием методов машинного обучения, Сеченовский университет. 2023

Original article

Classifier of bacterial and viral inflammation by hemogram analysis using machine learning methods

Andrey Andreevich Venerin[1]✉, Nikolay Igorevich Kanevskiy[2]

[12]N.V. Sklifosovskiy Institute of Clinical Medicine, Sechenov University, Moscow, Russian Federation

[12]Institute of Digital Medicine, Sechenov University, Moscow, Russian Federation

[1]venerin.andrey@gmail.com, http://orcid.org/0000-0002-8960-5772

[2]k_okin@mail.ru, http://orcid.org/0000-0003-4322-0110

Annotation. This paper presents a variant of a simplified hematological classifier grouping objects into normal, bacterial and viral inflammation groups. The paper presents implementations of machine learning methods with a teacher: a solver tree and a random forest. The results obtained can form the basis for writing a full-fledged classifier-assistant physician, which significantly speeds up the process of analyzing hemograms of patients.

Keywords: machine learning, hemograms, inflammation, random forest classifier, decision tree classifier

For citation: Venerin A.A., Kanevskiy N.I., Classifier of bacterial and viral inflammation by hemogram analysis using machine learning methods, Sechenov University. 2023

Введение. Вирусные и бактериальные инфекции до сих пор являются одними из лидирующих причин смерти и летальных осложнений, а также выступают экономическим бременем для общества во всем мире [1]. В связи с чем раннее выявление и правильное лечение имеют первостепенную важность. Одним из ключевых методов диагностики является общий анализ крови, позволяющий сразу же предположить этиологию заболевания. В веке цифровых технологий на помощь врачам приходит машинное обучение, которое в обозримом будущем сможет обеспечить практически моментальную интерпретацию гемограмм, а возможно и корректное назначение терапии, что значительно снизит вероятность возникновения осложнений [2]. Дифференциальная диагностика бактериального и вирусного воспаления имеет фундаментальное значение, так как этиотропная терапия имеет различные точки приложения, но порой вызывает затруднения ввиду схожести клинической картины [3]. Подобное часто наблюдается при дифференциальной диагностике пневмоний вирусного и бактериального генеза [4]. Необоснованное назначение антибактериальных препаратов при вирусных инфекциях может способствовать формированию антибиотикорезистентной флоры, вызывающий в дальнейшем продолжительные заболевания тяжелого течения и с неблагоприятным прогнозом, что мы сейчас и наблюдаем после массового применения антибиотиков для лечений COVID-ассоциированных пневмоний. При дифференциальной диагностике можно изучать экспрессию различных генов, но с экономической точки зрения на первом этапе это будет необоснованно [5]. В связи с этим оптимальным скрининговым методом является

общий анализ крови, так как при отклонении ключевых параметров от референса можно предположить этиологию заболевания и начать эмпирическое лечение, с целью скорейшего облечения состояния пациента.

Материалы и методы. Данные для анализа были получены в результате синтеза объектов с помощью библиотеки numpy языка программирования Python. Были созданы три первичных блока данных: норма, бактериальное воспаление, вирусное воспаление. Количество объектов каждого равно 1000 (n=1000). Далее данные были соединены конкатенацией из библиотеки pandas и перемешаны с помощью метода shuffle из библиотеки sklearn.utils. Целевой признак столбца norm конечного датасета был подготовлен к анализу числовым кодированием: {норма: 1, бактериальное воспаление: 2, вирусное воспаление: 3}. Итоговый датасет data содержит 3000 строк и 19 столбцов. Столбцы датасета: гемоглобин, эритроциты, лейкоциты, абс. сегментоядерные нейтрофилы, абс. палочкоядерные нейтрофилы, % сегментоядерные нейтрофилы, % палочкоядерные нейтрофилы, абс. нейтрофилов, % лимфоцитов, абс. лимфоцитов, % эозинофилов, абс. эозинофилов, % базофилов, абс. базофилов, % моноцитов, абс. моноцитов, тромбоциты, СОЭ, целевой категориальный признак. Среда разработки: colab. Язык разработки: Python 3.11. Основные библиотеки: pandas, numpy, sklearn. Использованные методы машинного обучения: DecisionTreeClassifier, RandomForestClassifier. Логистическая регрессия не использовалась в работе, поскольку целевой признак был составлен с помощью числового кодирования данных, несовместимого с регрессионными моделями. В качества основной метрики была использована f1-мера – взвешенное среднее гармоническое точности и полноты, чтобы не останавливаться отдельно на метриках precision и recall. Проверка моделей на адекватность проводилась с использованием DummyClassifier, прогнозирующего признак случайным образом. Во всех моделях random_state был равен 1. С кодом можно ознакомиться в открытом репозитории по адресу: https://github.com/AAVenerin.

Результаты. Подготовленный датасет не содержал пропусков, явных и неявных дубликатов, дисбаланса признаком и мультиколлинеарности в данных. При работе решающего дерева был подобран гиперпараметр max_depth=10 при range(1, 21, 1). F1-мера при использовании данного гиперпараметра равнялась 0.974, что является отличным показателем для классификатора. При работе случайного леса были подобраны гиперпараметры max_depth=5 и n_estimators=4. F1-мера при использовании данных гиперпараметров равнялась 0.979, что незначительно лучше результатов работы решающего дерева. При работе случайного классификатора использовалась стратегия ‘uniform’. F1-мера при использовании данного гиперпараметра равнялась 0.317, что

значительно хуже показателей работы выбранных моделей обучения. Таким образом, нами была обучена модель случайного леса, практически безошибочно классифицирующая объекта тестовой выборки после обучения и валидации.

Обсуждение. Продолжение работы требует внесения некоторых корректировок. Во-первых, в работе использовались синтетические данные. Несмотря на осторожность в генерировании датасетов, добавление реальных данных из биобанков было бы целесообразно. Также в синтетических данных не учитывается пол и возраст, поскольку перед нами стояла задача концептуального решения поставленной задачи без необходимых в будущем, но мешающих на этапе пилотного тестирования дополнительных параметров. Во-вторых, классификация осуществляется в достаточно небольшом приближении. Дифференцирование анализа, например, по наличию или отсутствию анемии, ее степени, добавление в анализ биохимического анализа крови – все это приблизило бы разработку к практическому внедрению в качестве программного обеспечения амбулаторных учреждений здравоохранения и лабораторий. Наконец, классификация – лишь первый шаг. Следует продолжать разработку алгоритмов автоматического назначения необходимых дополнительных лабораторных и инструментальных исследований, а также врачебных консультаций, основывающихся на результатах машинного анализа гемограмм пациентов.

Заключение. Нами была проделана работа по созданию простого классификатора гемограмм. В работе успешно себя показали методы DecisionTreeClassifier и RandomForestClassifier из библиотеки sklearn языка программирования Python. Развитие работы должно учитывать все дискуссионные аспекты, указанные в обсуждении к этой статье, что существенно ускорит внедрение данной и подобных методик в практической профилактическое здравоохранения.

Список источников:

1. Adam H. Agbaria, Guy Beck Rosen, Itshak Lapidot, Daniel H. Rich, Mahmoud Huleihel, Shaul Mordechai, Ahmad Salman, and Joseph Kapelushnik, Differential Diagnosis of the Etiologies of Bacterial and Viral Infections Using Infrared Microscopy of Peripheral Human Blood Samples and Multivariate Analysis, Analytical Chemistry 2018 90 (13), 7888-7895, DOI: 10.1021/acs.analchem.8b00017
2. Gunčar, G., Kukar, M., Notar, M. et al. An application of machine learning to haematological diagnosis. Sci Rep 8, 411 (2018). https://doi.org/10.1038/s41598-017-18564-8
3. Largman-Chalamish M, Wasserman A, Silberman A, Levinson T, Ritter O, Berliner S, et al. (2022) Differentiating between bacterial and viral infections by estimated CRP velocity. PLoS ONE 17(12): e0277401. https://doi.org/10.1371/journal. pone.0277401

4. Thomas J, Pociute A, Kevalas R, Malinauskas M, Jankauskaite L. Blood biomarkers differentiating viral versus bacterial pneumonia aetiology: a literature review. Ital J Pediatr. 2020 Jan 9;46(1):4. doi: 10.1186/s13052-020-0770-3. PMID: 31918745; PMCID: PMC6953310.
5. Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. Sci Transl Med. 2016 Jul 6;8(346):346ra91. doi: 10.1126/scitranslmed.aaf7165. PMID: 27384347; PMCID: PMC5348917.

References

1. Adam H. Agbaria, Guy Beck Rosen, Itshak Lapidot, Daniel H. Rich, Mahmoud Huleihel, Shaul Mordechai, Ahmad Salman, and Joseph Kapelushnik, Differential Diagnosis of the Etiologies of Bacterial and Viral Infections Using Infrared Microscopy of Peripheral Human Blood Samples and Multivariate Analysis, Analytical Chemistry 2018 90 (13), 7888-7895, DOI: 10.1021/acs.analchem.8b00017
2. Gunčar, G., Kukar, M., Notar, M. et al. An application of machine learning to haematological diagnosis. Sci Rep 8, 411 (2018). https://doi.org/10.1038/s41598-017-18564-8
3. Largman-Chalamish M, Wasserman A, Silberman A, Levinson T, Ritter O, Berliner S, et al. (2022) Differentiating between bacterial and viral infections by estimated CRP velocity. PLoS ONE 17(12): e0277401. https://doi.org/10.1371/journal. pone.0277401
4. Thomas J, Pociute A, Kevalas R, Malinauskas M, Jankauskaite L. Blood biomarkers differentiating viral versus bacterial pneumonia aetiology: a literature review. Ital J Pediatr. 2020 Jan 9;46(1):4. doi: 10.1186/s13052-020-0770-3. PMID: 31918745; PMCID: PMC6953310.
5. Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. Sci Transl Med. 2016 Jul 6;8(346):346ra91. doi: 10.1126/scitranslmed.aaf7165. PMID: 27384347; PMCID: PMC5348917.

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации.
Авторы заявляют об отсутствии конфликта интересов.
Contribution of the authors: the authors contributed equally to this article.
The authors declare no conflicts of interests.