ЭКСПЕРИМЕНТАЛЬНАЯ СТАТЬЯ

Бактериология, 2022, том 7, №2, с. 34–38
Bacteriology, 2022, volume 7, No 2, p. 34–38

DOI: 10.20953/2500-1027-2022-2-34-38

Определение эффективности алгоритмов библиотеки машинного обучения PyCaret для видовой идентификации микроорганизмов на основании данных времяпролетной масс-спектрометрии

М.Д.Хамитова[1], К.В.Детушев[2], А.Г.Богун[2]

[1]Автономная некоммерческая организация дополнительного профессионального образования «Школа анализа данных», Москва, Российская Федерация;
[2]ФБУН «Государственный научный центр прикладной микробиологии и биотехнологии» Роспотребнадзора, Оболенск, Московская область, Российская Федерация

Работа посвящена оценке возможности применения технологий машинного обучения для анализа данных, полученных при масс-спектрометрии бактериальных культур. В работе использовали масс-листы микроорганизмов видов *Escherichia coli* и *Staphylococcus aureus*, полученные с помощью MALDI-TOF масс-спектрометрии. Полученные данные обрабатывали с применением библиотеки PyCaret и выбранного в ходе исследования алгоритма Light Gradient Boosting Machine. Показано, что при анализе данных, полученных при идентификации штаммов бактерий *E. coli* и *S. aureus* из библиотеки PyCaret, алгоритм Light Gradient Boosting Machine позволяет получать результаты со степенью достоверности около 98%.
Ключевые слова: машинное обучение, масс-спектрометрия, MALDI-TOF, PyCaret

Для цитирования: Хамитова М.Д., Детушев К.В., Богун А.Г. Определение эффективности алгоритмов библиотеки машинного обучения PyCaret для видовой идентификации микроорганизмов на основании данных времяпролетной масс-спектрометрии. Бактериология. 2022; 7(2): 34–38. DOI: 10.20953/2500-1027-2022-2-34-38

Determining the effectiveness of the algorithms of the PyCaret machine learning library for species identification of microorganisms based on time-of-flight mass spectrometry data

M.D.Khamitova[1], K.V.Detushev[2], A.G.Bogun[2]

[1]Autonomous Non-Commercial Organization of Additional Professional Education "School of Data Analysis", Moscow, Russian Federation;
[2]State Research Center for Applied Microbiology and Biotechnology of Rospotrebnadzor, Obolensk, Moscow Region, Russian Federation

The work is devoted to assessing the possibility of using machine learning technologies for the analysis of data obtained by mass spectrometry of bacterial cultures. We used mass lists of microorganisms of *Escherichia coli* and *Staphylococcus aureus* species obtained using MALDI-TOF mass spectrometry. The obtained data were processed using the PyCaret library and the Light Gradient Boosting Machine algorithm selected during the study. It has been shown that when analyzing data obtained during the identification of bacterial strains of *E. coli* and *S. aureus* from the PyCaret library, the Light Gradient Boosting Machine algorithm makes it possible to obtain results with a confidence level of about 98%.
Key words: machine learning, mass spectrometry, MALDI-TOF, PyCaret

For citation: Khamitova M.D., Detushev K.V., Bogun A.G. Determining the effectiveness of the algorithms of the PyCaret machine learning library for species identification of microorganisms based on time-of-flight mass spectrometry data. Bacteriology. 2022; 7(2): 34–38. (In Russian). DOI: 10.20953/2500-1027-2022-2-34-38

Для корреспонденции:

Богун Александр Геннадьевич, кандидат биологических наук, ведущий научный сотрудник отдела коллекционных культур ФБУН «Государственный научный центр прикладной микробиологии и биотехнологии» Роспотребнадзора

Адрес: 142279, Московская область, г.о. Серпухов, р.п. Оболенск, Территория «Квартал А», 24
Телефон: (4967) 36-0000
E-mail: bogun@obolensk.org

Статья поступила 26.06.2022 г., принята к печати 30.06.2022 г.

For correspondence:

Aleksandr G. Bogun, PhD (Biological Sciences), Leading Researcher of the Department of Collection Cultures, State Research Center for Applied Microbiology and Biotechnology of Rospotrebnadzor

Address: SRCAMB 24 “Quarter A” Territory, Obolensk, 142279, Russian Federation
Phone: (4967) 36-0000
E-mail: bogun@obolensk.org

The article was received 26.06.2022, accepted for publication 30.06.2022

Технологии машинного обучения востребованы во многих областях исследований, связанных с обработкой данных. В данной работе нами была изучена возможность использования технологий машинного обучения для выявления видовой принадлежности штаммов микроорганизмов при идентификации с помощью времяпролетной масс-спектрометрии.

Технологии масс-спектрометрии получили широкое распространение для проведения быстрой видовой идентификации микроорганизмов. Отличительной особенностью этого подхода является универсальность процедуры, связанной с проведением анализа, поскольку для идентификации различных микроорганизмов может быть использован один аппаратно-программный комплекс, состоящий из времяпролетного масс-спектрометра с матрично-активированной лазерной десорбцией/ионизацией и ЭВМ. Другим важным преимуществом является возможность получать близкие данные с использованием разных моделей масс-спектрометров, в т.ч. функционирующих в разных лабораториях.

Ключевой составляющей аппаратно-программного комплекса является база данных, содержащая масс-спектры, характерные для изучаемых микроорганизмов. В настоящее время каждый производитель систем для идентификации бактерий, помимо оборудования собственного производства, создает программное обеспечение, включающее в себя базу данных и алгоритм анализа, которые не предназначены для использования с другим оборудованием.

В настоящее время особый интерес представляет создание объединенной, или универсальной, базы данных и алгоритма анализа, пригодных для обработки информации, получаемой при масс-спектрометрическом анализе на разных моделях оборудования. В связи с тем, что создание и проверка эффективности функционирования подобной базы данных и алгоритмов представляет собой сложную задачу, очень привлекательным является использование технологий машинного обучения. Существенным преимуществом технологий машинного обучения также является возможность выявлять ошибки, допускаемые при работе оператора.

Цель наших исследований заключается в оценке возможности использования технологий машинного обучения с использованием библиотеки PyCaret для идентификации микроорганизмов по результатам их исследования с помощью технологии MALDI-TOF масс-спектрометрии на примере *Escherichia coli* и *Staphylococcus aureus*.

Материалы и методы

Исходные данные. В работе были использованы результаты масс-спектрометрии (масс-листы), выбранные на основании идентификации с использованием комплекса MALDI-Biotyper (Bruker Daltonics) [1]. В табл. 1 приведен пример масс-листа, полученный в результате масс-спектрометрического исследования штамма *E. coli*. Масс-листы, использованные для машинного обучения в данной работе, могут быть предоставлены авторами работы по запросу.

В работе были использованы по 10 масс-листов для микроорганизмов каждого вида. Для машинного обучения были выбраны масс-листы, которые при анализе на системе MALDI-Biotyper показывали идентификацию до вида с уровнем индекса достоверности (Score Value) 2,3 и более. Максимальный показатель уровня достоверности идентификации (Score Value), получаемых на системе MALDI-Biotyper, равняется 3. В табл. 2 представлены показатели индекса достоверности для масс-листов, выбранных для работы на основании идентификации на системе MALDI-Biotyper.

Программы и способы машинного обучения. Для обучения модели потребовалось объединить в один массив данных исходные 10 масс-листов. Объединение осуществлялось с помощью программы flexAnalysis (Bruker Daltonics), данные экспортировались в Microsoft Office Excel.

Для машинного обучения была использована виртуальная машина, предоставленная сервером Google Colab (сервис, позволяющий писать код и работать с данными прямо в браузере, без установки программного обеспечения на компьютер).

Для машинного обучения мы используем библиотеку PyCaret, которая содержит 16 вариантов классификаторов, из которых мы сможем выбрать наиболее точно определяющий вид микроорганизмов [2].

PyCaret – библиотека машинного обучения с открытым исходным кодом на Python для обучения и развертывания моделей с учителем и без учителя в low-code среде. Low-code – это концепция создания информационных систем с помощью графических интерфейсов с минимальным (low-code) использованием ручного написания кода или вообще без него [3].

Целью применения концепции low-code является сокращение объема традиционного ручного кодирования и ускорение разработки бизнес-приложений. Кроме этого, важным преимуществом подхода является то, что вклад в проектирование может внести широкий круг людей, обладающих знаниями предметной области и понимающих бизнес-логику, а не только программисты [4].

PyCaret позволит пройти путь от подготовки данных до развертывания модели за несколько секунд. По сравнению с другими открытыми библиотеками машинного обучения PyCaret – это low-code альтернатива, которая поможет заменить сотни строк кода всего парой слов. Скорость проведения более эффективных экспериментов возрастет экспоненциально. PyCaret – это, по сути, оболочка Python над несколькими библиотеками машинного обучения, такими как scikit-learn, XGBoost, Microsoft LightGBM, spaCy и несколькими другими [3].

Результаты и обсуждение

В процессе использования платформы PyCaret нами были оценены 16 алгоритмов машинного обучения. В табл. 3 указана точность предсказаний. Наилучший результат обучения показал алгоритм Light Gradient Boosting Machine.

Light GBM – это быстрая, распределенная, высокопроизводительная структура повышения градиента, основанная на алгоритме дерева решений, используемая для ранжирования, классификации и многих других задач машинного обучения. Поскольку он основан на алгоритмах дерева ре-

Таблица 1. **Масс-лист, полученный при масс-спектрометрическом исследовании штамма *E. coli* с использованием времяпролетного масс-спектрометра microflex LRF (Bruker Daltonics)**

Соотношение заряд/масса	Время	Интенсивность	Соотношение сигнала к шуму	Разрешение	Площадь	Относительная интенсивность	Полуширина сигнала
2034,662	26859,37	825,000	5,220	427,267	4669,499	0,183	4,762
2066,516	27064,77	1285,000	8,130	436,424	8266,051	0,284	4,735
2082,189	27165,25	871,000	5,511	414,982	7176,067	0,193	5,018
2165,933	27695,85	940,000	6,039	457,247	6876,912	0,208	4,737
2179,899	27783,33	1479,000	9,635	411,198	11077,184	0,327	5,301
2407,230	29169,74	1063,500	7,007	463,096	6294,013	0,235	5,198
2554,516	30033,20	889,000	5,857	500,909	5798,413	0,197	5,100
2688,818	30799,05	1794,000	11,923	573,940	12397,909	0,397	4,685
2833,789	31604,55	1591,000	10,675	507,000	11011,353	0,352	5,589
3126,589	33170,80	1985,000	13,405	548,960	18004,459	0,439	5,695
3156,737	33327,82	1630,000	11,008	591,570	14588,484	0,361	5,336
3578,685	35451,29	1681,500	12,085	573,749	14501,007	0,372	6,237
3636,908	35734,23	1467,500	10,827	559,340	14181,474	0,325	6,502
3935,442	37150,62	1463,000	11,538	520,018	17634,956	0,324	7,568
4184,408	38291,16	1139,000	9,586	619,677	11914,554	0,252	6,753
4349,944	39030,77	620,000	5,272	602,390	6511,540	0,137	7,221
4365,103	39097,80	4247,000	36,111	586,463	42713,776	0,940	7,443
4438,059	39418,73	1447,000	12,304	706,106	16047,805	0,320	6,285
4448,261	39463,40	813,000	6,913	731,742	9259,544	0,180	6,079
4496,805	39675,24	1469,000	12,703	589,124	22349,111	0,325	7,633
4612,944	40177,45	1259,000	11,367	673,385	14203,693	0,279	6,850
4768,378	40839,77	2156,500	19,974	800,376	20312,114	0,477	5,958
4776,947	40875,97	2537,500	23,503	715,938	30253,346	0,562	6,672
5069,130	42091,30	983,000	9,197	761,461	10907,449	0,218	6,657
5097,053	42205,58	1345,000	12,583	677,223	15841,652	0,298	7,526
5149,946	42421,21	1919,000	18,470	610,545	34795,083	0,425	8,435
5367,691	43297,43	742,000	7,206	901,640	8036,934	0,164	5,953
5381,575	43352,69	4519,000	43,889	667,653	52350,680	1,000	8,060
6255,380	46697,40	4460,000	53,946	723,871	71261,788	0,987	8,642
6300,080	46861,99	955,000	11,557	726,175	18699,522	0,211	8,676
6315,619	46919,07	3025,000	36,607	712,466	49012,754	0,669	8,864
6411,633	47270,22	905,000	10,952	742,352	13290,284	0,200	8,637
6507,362	47617,71	721,000	8,760	596,306	13324,917	0,160	10,913
7158,042	49914,73	2625,500	34,156	758,046	42288,571	0,581	9,443
7273,756	50312,10	2358,500	30,771	771,325	40792,820	0,522	9,430
7870,209	52311,86	1623,500	23,012	709,803	41798,459	0,359	11,088
8367,506	53921,79	886,000	14,627	800,417	16204,169	0,196	10,454
8873,415	55511,12	920,000	16,867	874,169	14792,718	0,204	10,151
8991,265	55874,77	916,000	17,202	826,572	24022,826	0,203	10,878
9223,755	56585,24	911,000	18,390	897,472	14722,187	0,202	10,277
9533,194	57517,07	1977,000	40,093	928,315	45567,415	0,437	10,269
9550,708	57569,36	1570,000	32,842	1004,259	25787,815	0,347	9,510
10295,878	59751,02	858,000	19,461	908,734	15222,151	0,190	11,330

В масс-листах содержатся следующая информация: соотношение массы к заряду (m/z); время (time); интенсивность (Intens.); соотношение сигнала к шуму (SN); разрешение (Res.); площадь (Area); относительная интенсивность % (Rel. Intens.); полуширина сигнала (FWHM).

Таблица 2. **Показатели достоверности идентификации микроорганизмов (MALDI-Biotyper)**

№	Вид микроорганизма	Score Value	Дата эксперимента
1	*E. coli*	2,487	17.03.2022
2	*E. coli*	2,464	17.03.2022
3	*E. coli*	2,419	17.03.2022
4	*E. coli*	2,522	17.03.2022
5	*E. coli*	2,514	17.03.2022
6	*E. coli*	2,482	17.03.2022
7	*E. coli*	2,416	17.03.2022
8	*E. coli*	2,346	17.03.2022
9	*E. coli*	2,372	17.03.2022
10	*E. coli*	2,399	17.03.2022
11	*S. aureus*	2,276	19.03.2021
12	*S. aureus*	2,389	25.03.2021
13	*S. aureus*	2,399	25.03.2021
14	*S. aureus*	2,341	25.03.2021
15	*S. aureus*	2,355	18.10.2021
16	*S. aureus*	2,359	18.10.2021
17	*S. aureus*	2,360	10.12.2021
18	*S. aureus*	2,361	10.12.2021
19	*S. aureus*	2,352	31.01.2022
20	*S. aureus*	2,341	31.01.2022

шений, он разделяет лист дерева с наилучшим соответствием, тогда как другие алгоритмы повышения делят дерево по глубине или уровню, а не по листу. Таким образом, при выращивании на одном и том же листе в Light GBM листовой алгоритм может снизить потери по сравнению с поуровневым и, следовательно, приводит к гораздо лучшей точности, что редко может быть достигнуто любым из существующих алгоритмов повышения [5].

Отдельные деревья решений можно легко интерпретировать, просто визуализируя их структуру. Поскольку в модели градиентного бустинга содержатся сотни деревьев, ее нелегко интерпретировать путем визуализации входящих в нее деревьев. При этом хотелось бы, как минимум, понимать, какие именно признаки данных оказывают наибольшее влияние на предсказание композиции (рис. 1).

Можно сделать следующее наблюдение: признаки, используемые в верхней части дерева, влияют на окончательное предсказание для большей доли обучающих объектов, чем признаки, попавшие на более глубокие уровни. Таким образом, ожидаемая доля обучающих объектов, для которых происходило ветвление по данному признаку, может быть использована в качестве оценки его относительной важности для итогового предсказания. Усредняя полученные оценки важности признаков по всем решающим деревьям из ансамбля, можно уменьшить дисперсию такой оценки и использовать ее для отбора признаков.

Для определения важности параметров и исключения из данных параметров, которые не влияют на точность предсказания, нами был использован инструмент feature_importance (важность признаков). Метод оценки важности

Таблица 3. **Результаты использования различных алгоритмов машинного обучения**

Модель	Точность	
lightgbm	Light Gradient Boosting Machine	0,9928
catboost	CatBoost Classifier	0,9917
rf	Random Forest Classifier	0,9900
xgboost	Extreme Gradient Boosting	0,9895
dt	Decision Tree Classifier	0,9878
et	Extra Trees Classifier	0,9845
gbc	Gradient Boosting Classifier	0,9767
ada	Ada Boost Classifier	0,9396
ida	Linear Discriminant Analysis	0,8042
ridge	Ridge Classifier	0,8036
knn	K Neighbors Classifier	0,7582
qda	Quadratic Discriminant Analysis	0,6922
lr	Logistic Regression	0,6023
nb	Naïve Bayes	0,6007
dummy	Dummy Classifier	0,5641
svm	SVM-Linear Kernel	0,4965

feature_importance встроен в алгоритмы построения ансамблей деревьев и основан на вычислении суммарного уменьшения минимизируемого функционала ошибки с помощью ветвлений по рассматриваемому признаку. Сравниваемые пары листьев имеют разные значения разделения в узле на пути к этим листьям. Если условие разделения выполнено, объект переходит в левое поддерево, в противном случае он переходит в правое [6]. Согласно результатам оценки параметров (рис. 2) было принято решения исключить из данных параметры: площадь (Area) и полуширина сигнала (FWHM), время (time), разрешение (Res.), так как они имеют наименьшую важность.

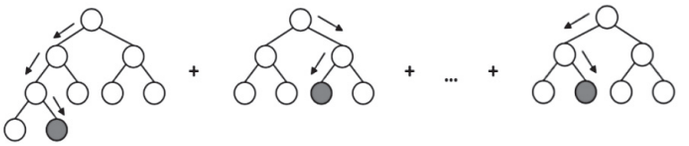

Рис. 1. **Разделение листов в дереве решений в алгоритме Light GBM.**

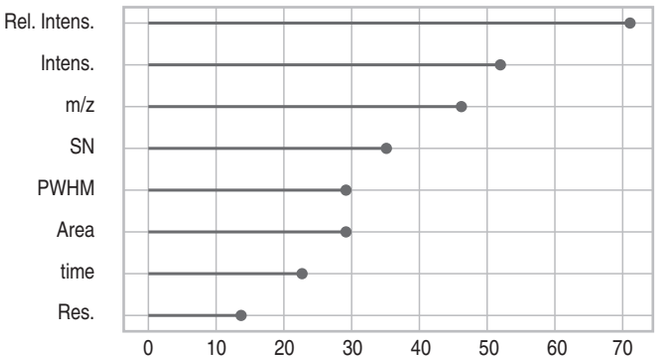

Рис. 2. **Оценка параметров, влияющих на точность определения вида.**

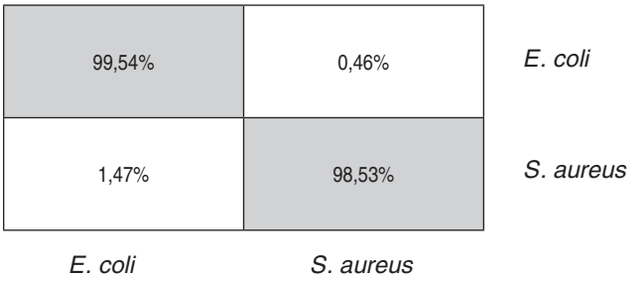

Рис. 3. **Матрица точности идентификации бактерий на тестовых данных.**

В процессе машинного обучения признаки, имеющие невысокую важность, создают «шумы», ухудшающие итоговые результаты предсказания.

Матрица показывает, что после обучения на тестовых данных алгоритм определил вид *E. coli* с точностью 99,54%, *S. aureus* – с точностью 98,53% (рис. 3).

Выводы

В проведенной нами работе показано, что технологии машинного обучения с использованием библиотеки PyCaret могут различать данные, полученные при рутинной идентификации патогенных микроорганизмов, методом масс-спектрометрии. Установлено, что при анализе данных, полученных при идентификации штаммов бактерий *E. coli* и *S. aureus*, алгоритм Light Gradient Boosting Machine позволяет получать наиболее достоверные результаты.

Полученные результаты позволяют сделать вывод, что использование технологий машинного обучения при идентификации микроорганизмов является перспективным направлением исследований.

Информация о финансировании

Работа выполнена в рамках отраслевой программы Роспотребнадзора.

Financial support

The work was carried out within the framework of the branch program of Rospotrebnadzor.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов, требующего раскрытия в данной статье.

Conflict of interest

Authors declare no conflict of interest requiring disclosure in this article.

Литература

1. Neville SA, Lecordier A, Ziochos H, Chater MJ, Gosbell IB, Maley MW, van Hal SJ. Utility of matrix-assisted laser desorption ionization-time of flight mass spectrometry following introduction for routine laboratory bacterial identification. J Clin Microbiol. 2011 Aug;49(8):2980-4. DOI: 10.1128/JCM.00431-11
2. Train – PyCaret Official [Электронный ресурс]. URL: https://pycaret.gitbook.io/docs/get-started/functions/train (дата обращения 20.06.2022).
3. Представляем PyCaret: открытую low-code библиотеку машинного обучения на Python. Хабр. URL: https://habr.com/ru/company/otus/blog/497770/ (дата обращения 20.06.2022).
4. Малокодовая разработка (Low-code) Loginom Wiki [Электронный ресурс]. URL: https://wiki.loginom.ru/articles/low-code.html (дата обращения 20.06.2022).
5. Machine Learning Mystery. URL: https://www.machinelearningmastery.ru/lightgbm-vs-xgboost-which-algorithm-win-the-race-1ff7dd4917d/
6. Feature importances with a forest of trees – scikit-learn 1.1.1 documentation [Электронный ресурс]. URL: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html (дата обращения 20.06.2022).

References

1. Neville SA, Lecordier A, Ziochos H, Chater MJ, Gosbell IB, Maley MW, van Hal SJ. Utility of matrix-assisted laser desorption ionization-time of flight mass spectrometry following introduction for routine laboratory bacterial identification. J Clin Microbiol. 2011 Aug;49(8):2980-4. DOI: 10.1128/JCM.00431-11
2. Train – PyCaret Official. URL: https://pycaret.gitbook.io/docs/get-started/functions/train (accessed 20.06.2022).
3. PyCaret: an open low-code machine learning library in Python. Habr. URL: https://habr.com/ru/company/otus/blog/497770/ (accessed 20.06.2022). (In Russian).
4. Low-income development (Low-code) Loginom Wiki. URL: https://wiki.loginom.ru/articles/low-code.html (accessed 20.06.2022). (In Russian).
5. Machine Learning Mystery. URL: https://www.machinelearningmastery.ru/lightgbm-vs-xgboost-which-algorithm-win-the-race-1ff7dd4917d/
6. Feature importances with a forest of trees – scikit-learn 1.1.1 documentation [Электронный ресурс]. URL: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html (accessed 20.06.2022).

Информация об авторах:

Хамитова Мария Дмитриевна, студентка Автономной некоммерческой организации дополнительного профессионального образования «Школа анализа данных»

Детушев Константин Владимирович, младший научный сотрудник отдела коллекционных культур ФБУН «Государственный научный центр прикладной микробиологии и биотехнологии» Роспотребнадзора

Information about authors:

Maria D. Khamitova, student, Autonomous Non-Commercial Organization of Additional Professional Education "School of Data Analysis"

Konstantin V. Detushev, Junior Researcher of the Department of Collection Cultures, State Research Center for Applied Microbiology and Biotechnology of Rospotrebnadzor