Известия Алтайского государственного университета. 2024. № 1 (135). С. 126–131.
Izvestiya of Altai State University. 2024. No 1 (135). P. 126–131.

Научная статья
УДК 519.687:004.912
DOI: 10.14258/izvasu(2024)1-18

Автоматическая классификация генетических мутаций на основе методов машинного обучения

Ольга Николаевна Половикова[1], Анастасия Станиславовна Маничева[2], Вячеслав Вячеславович Ширяев[3]

[1]Алтайский государственный университет, Барнаул, Россия, ponOlgap@gmail.com
[2]Алтайский государственный университет, Барнаул, Россия, manichevaas@mc.asu.ru
[3]ООО «ИТ Сфера», Барнаул, Россия, asmuddi628@gmail.com

Original article

Automatic Classification of Genetic Mutations Based on Machine Learning Methods

Olga N. Polovikova[1], Anastasiia S. Manicheva[2], Vyacheslav V. Shiryaev[3]

[1]Altai State University, Barnaul, Russia, ponOlgap@gmail.com
[2]Altai State University, Barnaul, Russia, manichevaas@mc.asu.ru
[3]IT Sphere LLC, Barnaul, Russia, asmuddi628@gmail.com

В данной статье описывается проблема определения вида генетической мутации раковой опухоли после секвенирования ее генома. Проблематика решения относится к задачам многоклассовой классификации. В работе предложен подход определения классов мутаций на основе их текстового описания с помощью методов машинного обучения, относящихся к группе обучения с учителем. Исследование проводилось на примере набора данных по онкологическим заболеваниям на основе анализа мутаций генома в клетках опухоли. Набор данных включает вид гена, его мутацию, текстовое описание генной мутации и класс мутации. Число классов равно девяти. В соответствии со спецификой исходных данных проведено обоснование выбора методов предобработки и векторизации текста, которые необходимо применить перед использованием методов машинного обучения. Построены классификаторы текстовых данных на основе моделей: k-ближайших соседей, деревьев решений, байесовского классификатора, логистической регрессии. По результатам моделирования получены оценки метрик качества классификации. Показано, что для исходных данных наилучшей моделью классификации является логистическая регрессия, показавшая меньшее значение функции потерь.

This paper considers the problem of identifying the type of genetic mutation of a cancer tumor after sequencing its genome. The problem solution relates to multi-class classification problems. The paper proposes an approach for the identification of mutation classes based on their text description using supervised machine learning methods. The study is carried out using a data set on cancer diseases obtained by the analysis of genome mutations in tumor cells. The data set includes the gene type, its mutation, a text description of the gene mutation, and the mutation class (with nine different classes overall). The paper provides an analysis and justification of the optimal text preprocessing and vectorization methods suitable for the source data at the inputs of the machine learning methods. There are several text data classifiers based on the k-nearest neighbors, decision trees, Bayesian classifiers, and the logistic regression machine learning methods. The obtained classification performance metrics after many simulations reveal that the best method to perform classification and identification of mutation classes is the linear regression method with the lowest error rate.

Ключевые слова: генетические мутации, методы машинного обучения, классификация, кодирование текста, токенизация, векторизация, метрики качества обучения, логарифмическая функция потерь, подбор гиперпараметров модели.

Для цитирования: Половикова О.Н., Маничева А.С., Ширяев В.В. Автоматическая классификация генетических мутаций на основе методов машинного обучения // Известия Алтайского государственного университета. 2024. № 1 (135). С. 126–131. DOI: 10.14258/izvasu(2024)1-18.

Keywords: genetic mutations, machine learning methods, classification, text encoding, tokenization, vectorization, training quality metrics, logarithmic loss function, selection of model hyperparameters.

For citation: Polovikova O.N., Manicheva A.S., Shiryaev V.V. Automatic Classification of Genetic Mutations Based on Machine Learning Methods. *Izvestiya of Altai State University*. 2024. No 1 (135). P. 126–131. (In Russ.). DOI:10.14258/izvasu(2024)1-18.

Введение

Вот уже более 30 лет ученые занимаются прочтением генома человека. Исследования в этом направлении начались еще в конце XX столетия, а полная расшифровка была завершена в 2022 г. Тем не менее исследования по «прочтению» данных, которые удалось расшифровать, являются актуальными и на данный момент [1].

Ученые рассчитывали, что с полной расшифровкой генома произойдет существенный прорыв в области медицины и лечения человека. Однако здоровье человека определяется не только его геномом, но также и геномом тех организмов, которые с ним сосуществуют и которые причастны к его заболеваниям, например, бактерии и вирусы [2].

Важно также знать полиморфизмы (вариации) конкретных генов и как они влияют на здоровье человека. Что влечет за собой необходимость исследования генов целых популяций человека, собранных по этническому, национальному, территориальному или другим признакам. Сбором и исследованием геномов конкретных популяций занимаются различные ведомства, например, исландская компания «deCODE genetics» собрала генетическую информацию о двух третьих населения Исландии [3].

За последние несколько десятков лет тематика и методы точной или «прецизионной» медицины активно обсуждаются и используются на практике [4–9]. Основной потенциал прецизионного подхода в диагностике и лечении основывается на возможностях применения результатов генетического тестирования для составления индивидуального плана лечения каждого конкретного пациента.

Генетическое тестирование, а именно секвенирование генома опухолевой ДНК, является востребованной диагностикой, на результатах которой построены некоторые современные методы лечения. Секвенирование опухоли — это метод получения своего рода молекулярного сканирования ДНК, извлеченного из опухолевых клеток, полученных из образца биопсии, из крови или костного мозга пациента [6]. Эта информация предоставляет подробную информацию о том, какие области ДНК опухоли отличаются от ДНК неопухолевых клеток, а интерпретация данных геномного секвенирования дает представление о мутациях, которые могут являться онкологией. Секвенирование генома опухолевой ДНК может помочь идентифицировать первичный очаг опухоли и ключевые гены рака, тем самым скорректировать лечение. Кроме этого, результатом секвенирования является геномный профиль с выявленными мутациями [10]. На основе выявленных мутаций и накопленного банка данных раковых популяций можно спрогнозировать реакцию на конкретное лечение, а также определить тяжесть заболевания.

После секвенирования раковая опухоль может иметь тысячи генетических мутаций. Главная проблема заключается в определении вида мутации. На основе имеющихся данных нужно научиться отличать мутации, способствующие росту опухоли, от нейтральных мутаций. «Ручной» (не автоматизированный) способ определения вида мутации ресурсозатратен, а также требует привлечения специалистов.

Современные подходы и методы в области обработки, хранения и анализа данных предоставляют исследователям инструмент для расшифровки и интерпретации ДНК раковой опухоли. Решение подобных задач тесно связано с разработкой и применением алгоритмов интеллектуального анализа данных и математических моделей для обработки результатов геномного секвенирования [11].

Исходные данные

Материал для исследования получен с платформы для проведения соревнований по анализу данных Kaggle с применением методов машинного обучения и искусственного интеллекта [12]. Предоставленные данные являются реальными, полученными в результате лабораторных исследований и ручной маркировки специалистами.

Данные сформированы по автоматической классификации онкологических заболеваний на основе анализа мутаций генома в клетках опухоли. В настоящее время такая интерпретация генетических мутаций выполняется вручную, что сопряжено с большими трудозатратами и серьезными временными издерж-

ками. Ручная классификация выполняется клиническим патологоанатомом, который определяет класс, просматривая текстовую информацию из специализированной литературы.

Набор данных предоставляется двумя файлами. Первый файл предоставляет информацию о генетической мутации и классе, которому данная мутация принадлежит (ID, Gene, Variation, Class). Во втором файле содержится текстовое описание генетических мутаций (ID, Text). Данные первого и второго наборов связаны через уникальный идентификатор.

Выборка данных для обучения включает 3321 элемент. Согласно аннотированию данных, генетическая мутация может быть отнесена к девяти различным классам по 264 видам представленных генов. Данные являются несбалансированными по видам генома и классу мутации.

Информация в файлах объединяется в единый набор данных, где для каждого экземпляра имеется тип генома, класс мутации и текстовое описание. В целях применения алгоритмов машинного обучения текстовая информация векторизуется.

Выбор алгоритмов машинного обучения

Исходя из предварительного анализа полученных данных, решаемую задачу классификации генов следует отнести к задаче контролируемого машинного обучения, которая прогнозирует распределение экземпляров данных по нескольким классам (категориям).

После предварительного выбора алгоритмов классификации для каждого строится модель, на основе которой выполняется обучение (с использованием обучающего набора данных с метками). Результатом работы каждой такой модели является построение классификатора, который умеет прогнозировать класс для новых экземпляров, но уже без метки. Выбор одной или нескольких наиболее подходящих моделей осуществляется на результатах тестового набора данных.

Применение технологий, методов и алгоритмов искусственного интеллекта для классификации генетических мутаций в рамках исследования обусловливает преобразование текстового описания мутаций в числовое представление (с максимальной передачей смыслового содержания) с помощью алгоритмов и технологий обработки естественного языка (NLP: Natural Language Processing) [13–14].

Решение задачи в области компьютерного анализа и синтеза текстов на естественных языках представляется разбиением на несколько последовательных подзадач, каждая из которых выполняется отдельно и своим набором алгоритмов и методов: 1) предварительная обработка, разбиение текстового описания на токены и векторизация, 2) классификации мутаций генов.

Задачу векторизации токенов можно формализовать следующим образом: необходимо построить отображение (F), которое каждому элементу из множества текстов ставит в соответствие набор признаков: $F : T \to X$, где X — это признаковое пространство. Для цифрового представления естественно-языковых конструкций могут использоваться три группы подходов: 1) частотный подход, 2) тематическое моделирование, 3) дистрибутивная семантика.

Для исследования был выбран частотный подход на основе «мешка слов» с использованием меры TF–IDF, который является произведением двух значений частот:

$$tf - idf(t,d,D) = tf(t,d) \cdot idf(t,D),$$

где $tf(t,d)$ — частота слова, т.е. отношение числа вхождений некоторого слова к общему числу слов документа, показывает, насколько часто слово используется в данном конкретном документе; $idf(t,D)$ — обратная частота документа, т.е. это инверсия частоты, с которой данное слово встречается в документах всей коллекции, показывает, насколько нечасто данное слово встречается во всех рассматриваемых документах.

Обратная частота документа уменьшает вес слов, которые используются часто и являются некими «связками»: описывают больше предметную область в целом, чем отдельный класс. Получается, что уникальные слова, «привязанные» к конкретному классу, обладают большим значением меры TF–IDF. Метод TfidfVectorizer библиотеки sklearn языка программирования Python позволяет рассчитать необходимые весовые коэффициенты для определения меры TF–IDF.

Учитывая специфику решаемой целевой задачи — классификации, следует построить отображение G, которое ставит в соответствие каждому вектору признакового пространства метку или класс: $G : X \to \{1,2,...,9\}$, где 1,2,..,9 — идентификаторы классов.

Задача классификации мутаций генов на основе их текстового описания представляется двумя этапами: 1) построение пространства признаков (извлечение признаков), 2) разбиение признакового пространства на классы.

В работе рассматривались многоклассовые классификаторы на основе следующих моделей: *k-ближайших соседей*, *деревьев решений*, *байесовского классификатора*, *логистической регрессии*.

В основе *модели k-ближайших соседей* лежит гипотеза компактности: если метрика расстояния между объектами классификации определена адекватно, то схожие объекты чаще лежат в одном классе, чем в разных. Схожесть определяется по евклидову расстоянию.

Процесс построения *модели деревьев решений* заключается в последовательном, рекурсивном разбиении обучающего множества на подмножества с применением решающих правил вида «Если ..., то ...» в узлах. Процесс разбиения продолжается до тех пор, пока все узлы в конце всех ветвей не будут объявлены листьями.

Модель байесовского классификатора основана на применении теоремы Байеса и является вероятностным алгоритмом. Одним из основных преимуществ данного алгоритма является возможность применения при «небольшом» количестве данных для обучения.

В основе *модели логистической регрессии* лежит вычисление вероятности того, что некоторое значение принадлежит к определенному классу.

В качестве метрик, оценивающих качество обучения модели многоклассовой классификации, были выбрана *accuracy* (доля правильных ответов) и логарифмическая функция потерь *LogLoss* (среднее значение от логарифмов вероятностей классов назначения). Наиболее успешной будет считаться та модель, на которой *LogLoss* принимает меньшее значение.

Применение моделей машинного обучения для построения классификаторов

Получение векторного представления слов текстового описания осуществлялось на основе метода «мешка слов» с использованием меры TF–IDF, при этом текстовые категориальные признаки (Gene, Variation) также преобразованы в числовое представление с использованием целевого кодирования. Данный вид кодирования позволяет преобразовывать в числовые значения не ранжированные категориальные признаки и не изменяет размерность используемых данных обучения.

Для подбора оптимальных гиперпараметров моделей классификации был использован специальный метод RandomizedSearchCV, реализующий подбор комбинации параметров из заданных диапазонов значений, основываясь на выбранной метрике.

Результаты подбора гиперпараметров для моделей k-ближайших соседей, деревьев решений, байесовского классификатора, логистической регрессии и значения метрик качества представлены в таблице.

Метрики качества моделей классификации

Модель	*Accuracy*	*LogLoss*
k-ближайших соседей	0,519	1,469
Деревья решений	0,586	1,295
Байесовский классификатор	0,614	2,109
Логистическая регрессия	0,655	1,063

Минимальное значение функции потерь (LogLoss) получено при использовании модели логистической регрессии: LogLoss=1,063, при этом соответствующая точность — наибольшая из представленных: accuracy=0,655.

Заключение

Для решения задач классификации на основе текстового описания следует определиться с технологиями и подходом для кодирования исходных данных. Числовое кодирование является основным этапом предобработки текста для методов машинного обучения. Существует несколько методов, которые используются для многоклассовой классификации на основе текстового представления данных.

В процессе исследования были построены и оценены несколько моделей для определения класса генной мутации, чтобы выбрать оптимальную модель в рамках задачи классификации. Практической реализацией были оценены модели k-ближайших соседей, деревьев решений, байесовского классификатора, логистической регрессии и определена та, которая «лучше видит» закономерности между исходными данными и результирующим классом. Как метрика качества обучения в исследовании была выбрана функция логистических потерь LogLoss.

Из нескольких построенных моделей классификации текста наиболее успешной была признана модель логистической регрессии, на которой функция LogLoss принимает меньшее по сравнению с результатами других моделей значение.

Библиографический список

1. Код жизни: прочесть не значит понять // Kaggle. URL: https://biomolecula.ru/articles/kod-zhizni-prochest-neznachit-poniathttp://archive.expert.ru/expert/ (дата обращения: 10.11.2023).

2. Третья фаза ENCODE обнаружила тысячи новых взаимодействий внутри генома // PRC NEWS. URL: https://pcr.news/novosti/tretya-faza-encode-obnaruzhila-tysyachi-novykh-vzaimodeystviy-vnutri-genoma/ (дата обращения: 10.11.2023).

3. The Encyclopedia of DNA Elements (ENCODE) // National Human Genome Research Institute. URL: https://www.genome.gov/Funded-Pro-grams-Projects/ENCODE-Project-ENCyclopedia-Of-DNA-Elements (дата обращения: 10.11.2023).

4. The ENCODE Project Consortium et al. Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes // Nature. 2020. № 583. P. 699–710. DOI: 10.1038/s41586-020-2493-4

5. Vnencak-Jones C., Berger M., Pao W. Types of Molecular Tumor Testing // My Cancer Genome. URL: https://www.mycancergenome.org/content/molecular-medicine/types-of-molecular-tumor-testing/ (дата обращения: 10.11.2023).

6. Гаджиев Я., Шалбузова К. Применение методов машинного обучения в прогнозировании и раннем обнаружении рака // Sciences of Europe. 2022. № 108. С. 46–50.

7. Гусев А.В., Гаврилов Д.В., Корсаков И.Н. и др. Перспективы использования методов машинного обучения для предсказания сердечнососудистых заболеваний // Врач и информационные технологии. 2019. № 3. С. 41–47.

8. Гусев А.В., Новицкий Р.Э., Ившин А.А., Алексеев А.А. Машинное обучение на лабораторных данных для прогнозирования заболеваний // Фармакоэкономика. Современная фармакоэкономика и фармакоэпидемиология. 2021. № 4. С. 581–592. DOI: 10.17749/2070-4909/farmakoekonomika.2021.115

9. Раскина К.В., Мартынова Е.Ю., Перфильев А.В. и др. От персонализированной к точной медицине // Рациональная фармакотерапия в кардиологии. 2017. № 1. С. 69–79. DOI: 10.20996/1819-6446-2017-13-1-69-79

10. Emmert-Streib F. Personalized Medicine: Has it Started yet? A Reconstruction of the Early History // Front Genet. 2013. Vol. 3. № 313. DOI: 10.3389/fgene.2012.00313

11. 3 главных причины для геномного секвенирования рака // Блог сайта addon. URL: https://addon.life/ru/2021/08/02/genomic-sequencing-cancer/ (дата обращения: 10.11.2023).

12. Personalized Medicine: Redefining Cancer Treatment // Kaggle. URL: https://www.kaggle.com/competitions/msk-redefining-cancer-treatment/overview (дата обращения: 10.11.2023).

13. Обработка естественного языка // Машинное обучение. URL: https://www.dmitrymakarov.ru/intro/topic-identification-19/ (дата обращения: 10.11.2023).

14. Самигулин Т.Р., Джурабаев А.Э. Анализ тональности текста методами машинного обучения // Научный результат. Информационные технологии. 2021. № 1. С. 55–62. DOI: 10.18413/2518-1092-2021-6-1-0-7

References

1. Code of Life: Reading Does not Mean Understanding. *Kaggle*. URL: https://biomolecula.ru/articles/kod-zhizni-prochest-ne-znachit-poniathttp://archive.expert.ru/expert/ (accessed: 10.11.2023). (In Russ.).

2. Phase 3 of ENCODE Discovered Thousands of New Interactions within the Genome. *PRC NEWS*. URL: https://pcr.news/novosti/tretya-faza-encode-obnaruzhila-tysyachi-novykh-vzaimodeystviy-vnutri-genoma/ (accessed: 10.11.2023). (In Russ.).

3. The Encyclopedia of DNA Elements (ENCODE). *National Human Genome Research Institute*. URL: https://www.genome.gov/Funded-Pro-grams-Projects/ENCODE-Project-ENCyclopedia-Of-DNA-Elements (accessed: 10.11.2023).

4. The ENCODE Project Consortium et al. Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes. *Nature*. 2020. No 583. P. 699–710. DOI: 10.1038/s41586-020-2493-4

5. Vnencak-Jones C., Berger M., Pao W. Types of Molecular Tumor Testing. My Cancer Genome. URL: https://www.mycancergenome.org/content/molecular-medicine/types-of-molecular-tumor-testing/ (accessed: 10.11.2023).

6. Gadzhiev Ya., Shalbuzova K. Application of Machine Learning Methods in Cancer Prediction and Early Detection. *Sciences of Europe*. 2022. No 108. P. 46–50. (In Russ.)

7. Gusev A.V., Gavrilov D.V., Korsakov I.N. and etc. Prospects for Using Machine Learning Methods to Predict Cardiovascular Diseases. *Doctor and information technology*. 2019. No 3. P. 41–47. (In Russ.)

8. Gusev A.V., Novickij R.E., Ivshin A.A., Alekseev A.A. Machine Learning on Laboratory Data for Disease Prediction. Pharmacoeconomics. *Modern Pharmacoeconomics and Pharmacoepidemiology*. 2021. No 4. P. 581–592. DOI: 10.17749/2070-4909/farmakoekonomika.2021.115 (In Russ.)

9. Raskina K.V., Martynova E.Yu., Perfil'ev A.V. and etc. From Personalized to Precision Medicine. *Rational Pharmacotherapy in Cardiology*. 2017. No 1. P. 69–79. DOI: 10.20996/1819-6446-2017-13-1-69-79 (In Russ.)

10. Emmert-Streib F. Personalized Medicine: Has it Started yet? A Reconstruction of the Early History. *Front Genet*. 2013. Vol. 3. No 313. DOI: 10.3389/fgene.2012.00313

11. Top 3 Reasons for Genomic Sequencing of Cancer. *Site Blog Addon*. URL: https://addon.life/ru/2021/08/02/genomic-sequ-encing-cancer/ (accessed: 10.11.2023). (In Russ.)

12. Personalized Medicine: Redefining Cancer Treatment. *Kaggle*. URL: https://www.kaggle.com/competitions/msk-redefining-cancer-treatment/overview (accessed: 10.11.2023).

13. Natural Language Processing. *Machine Learning*. URL: https://www.dmitrymakarov.ru/intro/topic-identification-19/ (accessed: 10.11.2023). (In Russ.)

14. Samigulin T. R., Dzhurabaev A. E. Analysis of Text Sentiment Using Machine Learning Methods. Scientific Result. *Information Technology*. 2021. No 1. P. 55–62. (In Russ.). DOI: 10.18413/2518-1092-2021-6-1-0-7

Сведения об авторах

О.Н. Половикова, кандидат физико-математических наук, доцент, доцент кафедры информатики, Алтайский государственный университет, Барнаул, Россия;

А.С. Маничева, кандидат технических наук, доцент, доцент кафедры теоретической кибернетики и прикладной математики, Алтайский государственный университет, Барнаул, Россия;

В.В. Ширяев, программист отдела разработки систем визуализации данных, ООО «ИТ Сфера», Барнаул, Россия.

Information about the authors

O.N. Polovikova, Candidate of Sciences in Physics and Mathematics, Associate Professor, Associate Professor of the Department of Informatics, Altai State University, Barnaul, Russia;

A.S. Manicheva, Candidate of Sciences in Technology, Associate Professor, Associate Professor of the Department of Theoretical Cybernetics and Applied Mathematics, Altai State University, Barnaul, Russia

V.V. Shiryaev, Programmer of the Department of Data Visualization Systems Development, IT Sphere LLC, Barnaul, Russia.