Предсказание конфликтов при клинической классификации с помощью машинного обучения

К.А. Мусин[1], А.В. Гайдель[1,2]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086
[2]Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001

Аннотация. Клиническая классификация генетического варианта человека может привести к противоречивым классификациям. Наличие конфликтов определяется вручную лабораторными методами. Если есть конфликт, тогда трудно интерпретировать результат. В этой работе с помощью алгоритмов машинного обучения удалось обучить нейронную сеть прогнозированию конфликтов с точностью до 79%, а также определить, какие параметры являются наиболее важными в классификации.

1. Введение

В силу наличия связей определенных генов с различными морфологическими и физиологическими особенностями индивидуума, есть необходимость правильно их определить для постановки корректного диагноза пациенту [1]. Такой процесс именуется клинической классификацией. Зачастую возникают конфликты при возникновении противоречивых классификаций – это противодействует установлению корректной корреляции между геномом и болезнью. Для поиска таких конфликтов приходится задействовать лабораторные методы, которые являются весьма затратными, ибо требуют дополнительных исследований [2].

В данной работе проводится исследование возможности автоматизации процесса установления конфликтов с помощью метода машинного обучения Random forest и методов подготовки данных: Feature Hasher, One Hot Encode, Label Encoder, а также собственные методы, благодаря которым возможно интерпретировать данные для решения задачи.

Модели в работе обучались на данных, полученных из общедоступного медицинского архива Clinvar, в котором хранится информация о взаимосвязи между вариациями геномов и фенотипов с подтверждающими данными [2].

2. Подготовка данных для обучения модели

2.1. Формирование данных

Для работы с методами машинного обучения необходимо, чтобы все данные были представлены в численном виде. Были определены такие характеристики, данные в которых представлены не числами, а также количество уникальных значений в них с целью установления используемого метода для подготовки данных.

Характеристика PolyPhen (Polymorphism Phenotyping) состоит из нескольких описаний возможных влияний аминокислотного замещения на структуру и функцию человеческого

белка. Для такого набора данных был применен метод Label Encoder, присваивающий каждому состоянию определенное число.

Такая характеристика, как EXON (Экзон) – часть гена, кодирующая аминокислоты, была представлена в исходных данных, как часть экзонов от их общего числа в организме, что требовало дополнительной обработки с помощью методов работы со строками из библиотеки Python'a.

Для таких характеристик, как cDNA_position (положение пары генов в последовательности дополнительных ДНК), CDS_position (положение базовой пары генов в кодирующей области), Protein_position (положение аминокислоты в белке) данные были представлены в виде диапазонов положений, поэтому были подсчитаны их медианные значения, выражаемые скаляром.

Характеристики: REF (аллель сравнения), ALT (альтернативный аллель), CHROM (вариант хромосомы), Allele (аллель), Consequence (тип последствия) содержат количество различных значений от 24 до 866. Из-за этого применимым способом кодирования этих значений можно считать метод Feature hasher, векторизующий признаки в определенное количество столбцов, представимых в виде элементов булевы алгебры.

Такие характеристики, как: CLNVC (тип варианта), IMPACT (модификатор воздействия для типа последствия), BIOTYPE (биотип) имеют достаточно малый набор различных значений, поэтому для них хорошо применим метод One Hot Encode, создающий для каждого значения характеристики свой вектор. Также были проведены исследования по оставшимся столбцам, в некоторых из них значения отсутствовали, что лишь усложняло бы работу метода Random forest. В некоторых, как CLNHGVS (характеристика, содержащая описание уровня расположения генома) все значения были уникальными, отсюда следует, что корреляция между ними равна нулю, тогда и установление связей, играющих роль для классификатора невозможно. Поэтому часть информации была убрана.

2.2. Описание выбранного метода машинного обучения

Для данной задачи прогнозирования есть возможность воспользоваться методом бинарной классификации [3]. Подходящим методом является Random forest (Случайный лес), который относится к семейству ансамблевых методов, при это сам ансамбль состоит из простейших моделей: Decision tree (Дерево решений).

Дерево решений описывается следующим образом: пусть даны векторы обучения $x_i \in R^n, i = 1, \dots, l$ и вектор меток $y \in R^l$: дерево решений рекурсивно разделяет пространство так, что выборки с одинаковыми метками группируются вместе. Пусть данные в узле m будут представлены Q. Для каждого кандидата разделение $\theta = (j, t_m)$, состоящее из характеристики j и порога t_m, делит данные на подмножества $Q_{left}(\theta)$ и $Q_{right}(\theta)$, где $Q_{left}(\theta) = (x, y)|x_j \leq t_m$, $Q_{right}(\theta) = Q \setminus Q_{left}(\theta)$. Затем вычисляется функция $G(Q, \theta) = \frac{n_{left}}{N_m} H\left(Q_{left}(\theta)\right) + \frac{n_{right}}{N_m} H(Q_{right}(\theta))$, где $H()$ – это мера энтропии, задаваемая формулой:

$$H(X_m) = -\sum_k p_{mk} \log(p_{mk}).$$

Далее выбираются параметры по следующему критерию: $\theta^* = argmin_\theta G(Q, \theta)$. Подмножества $Q_{left}(\theta^*)$ и $Q_{right}(\theta^*)$ определяются до тех пор, пока не будет достигнута максимальная доступная глубина: $N_m < min_{samples}$ или $N_m = 1$ [4].

Взаимодействие решающих деревьев в алгоритме Случайный лес осуществляется с помощью подхода bagging – создание независимых моделей для оценки, а затем усреднение их прогнозов по следующей формуле:

$$S_l = \frac{1}{l} \sum_{l=1}^{L} w_l,$$

где L – число независимых базовых моделей, а w_l – полученный датасет каждой моделью [5].

3. Результат обучения моделей

Для определения наилучшей модели Random forest были проанализированы модели с различным количеством деревьев при разной глубине дерева. Сравнение происходило по F-мере, являющейся совместной оценкой точности и полноты, для классификации наличия конфликта, так как это является наиболее сложной задачей. Также было выяснено, что прогноз точнее при использовании автоматической глубины, когда дерево само решает, необходимо ли дальше искать новые связи или нет.

Были проведены исследования различных моделей Случайного леса, в итоге лучшим классификатором оказалась модель со 128 решающими деревьями, результат лучшего классификатора для тестовой выборки представлен на рисунке 1. Прогноз был повышен на 16% с помощью метода predict_proba(), понижением порога до 0.325.

```
Classification Report :
                precision    recall  f1-score   support

           0       0.86      0.73      0.79     12133
           1       0.46      0.67      0.55      4164

    accuracy                           0.72     16297
   macro avg       0.66      0.70      0.67     16297
weighted avg       0.76      0.72      0.73     16297
```

Рисунок 1. Результат прогнозирования классификатором для тестовой выборки.

Как можно видеть на рисунке 1, классификатор лучше справляется с определением ситуации, где отсутствует конфликт геномов. Для конфликтных ситуаций вероятность определения верных случаев ниже.

Далее на рисунке 2 приведен график ROC-кривой, позволяющий оценить качество бинарной классификации [6].

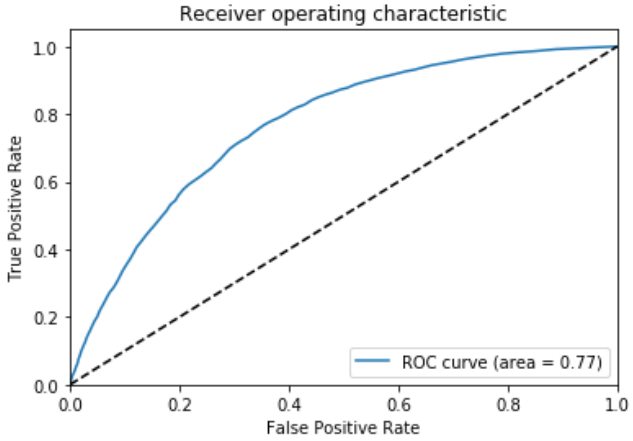

Рисунок 2. ROC-кривая для определения качества бинарной классификации.

Исходя из рисунка 2, площадь показания оценивается в 0.77, качество классификатора определяется тем, на сколько данный показатель высок.

4. Заключение

В данной работе было произведено исследование эффективности алгоритмов машинного обучения для задачи установления генетического конфликта при клиническом анализе. Полученный результат точности в 79% показывает, что применение машинных методов для задач генетической классификации и установления конфликтов является возможным, несмотря на сложность интерпретации исходных параметров для человека. Такое применение уменьшит

затраты на дополнительные медицинские исследования, а также позволит предсказать необходимость обследования для пациентов с подозрением на заболевание.

5. Благодарности
Работа выполнена при поддержке грантов РФФИ № 19-29-01235 мк и № 19-29-01135 мк, экспериментальные исследования – в рамках госзадания ИСОИ РАН – филиала ФНИЦ «Кристаллография и Фотоника» РАН (соглашение № 007-ГЗ/Ч3363/26).

6. Литература
[1] Мендель, Г.И. Опыты над растительными гибридами / Г.И. Мендель – М.: Наука, 1965. – 163 с.
[2] ClinVar [Electronic resource]. – Access mode: https://www.ncbi.nlm.nih.gov/clinvar/ (07.09.2019).
[3] Классификация [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Классификация (10.09.2019).
[4] Breiman, L. Classification and regression trees / L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone // International Biometric Society. – 1984. – Vol. 40(3). DOI: 10.2307/2530946.
[5] Decision trees [Electronic resource]. – Access mode: https://scikit-learn.org/stable/modules/tree.html#mathematical-formulation (23.09.2019).
[6] Receiver operaring characteristic [Electronic resouce]. – Access mode: https://en.wikipedia.org/wiki/Receiver_operating_characteristic (01.10.2019).

Machine learning algorithms in the prediction of conflicts in clinical classification of genetic variants

K.A. Musin[1], A.V. Gaidel[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

Abstract. The clinical classification of a person’s genetic variant can lead to conflicting classifications. The presence of conflicts is determined manually by laboratory methods. If there is a conflict, then there is a difficulty in interpreting the result. In this work, with the help of machine learning algorithms, it was possible to train the neural network to predict conflicts with an accuracy of 77%, and also to determine which parameters are most important in classification.