ПРОГНОЗИРОВАНИЕ РЕГУЛЯТОРНЫХ МИШЕНЕЙ ПРОКАРИОТИЧЕСКОЙ РНК С ПОМОЩЬЮ ГЛУБОКОГО МАШИННОГО ОБУЧЕНИЯ

М.А. Сиколенко

Белорусский государственный университет, Минск, Беларусь
E-mail: mishasikolenko@mail.ru

Малые регуляторные РНК широко распространены у прокариотов, причем наиболее изученное семейство из этих не кодирующих генов соответствует транс-действующим регуляторам, которые связываются с их мишенями путем парного взаимодействия оснований. Учитывая увеличивающуюся частоту идентификации этих генов, важно, чтобы методы для выявления их регуляторных целей не отставали. В данной статье исследуются взаимодействия между малыми РНК и их целями с использованием подхода глубокого машинного обучения. Получена метрика качества AUC, равная 0,906. Модель в данной работе обучается на том же наборе данных, что и алгоритм TargetRNA3, и демонстрирует лучшее качество классификации по сравнению с последним.

Ключевые слова: малые регуляторные РНК; нейронные сети; машинное обучение; биоинформатика; глубокое машинное обучение.

АНАЛИЗ НАБОРА ДАННЫХ

Малые РНК широко распространены и играют важную роль в регуляции генов, действуя на посттранскрипционном уровне они связываются с целевыми мРНК, влияя на их трансляцию или стабильность. Они способны воздействовать на несколько мРНК, что позволяет им оказывать широкое влияние на клеточные процессы. Однако их аннотация может быть сложной из-за разнообразия размеров, функций и степени консервативности [2].

В рамках обучающей выборки был использован набор данных для обучения модели TargetRNA3 с целью проведения корректного сравнения эффективности двух моделей. Данный набор данных состоит из 118 колонок. Первые 5 – не численные признаки. В оставшихся 112 колонках содержится пространство признаков, которые были рассмотрены создателями модели TargetRNA3. Кроме того, в наборе данных присутствует целевая переменная «Evinced Interaction», представленная в бинарном формате, указывающая наличие, экспериментальное подтверждение, (единица) или отсутствие (ноль) взаимодействия между малыми РНК и их мишенями. Полный набор данных доступен в файле формата .CSV [1].

ВЫБОР ПРИЗНАКОВ

При работе с обширными наборами признаков необходим качественный отбор, так как модели машинного обучения могут испытывать затруднения в выявлении значимых характеристик. При создании модели TargetRNA3 применялись статистическое методы для отбора признаков, в то время как в данной работе применялись методы интеллектуального анализа данных. Изначально в наборе данных используется пространство признаков, включающее результаты работы других алгоритмов, однако для повышения независимости решений от выводов других программ было принято решение сосредоточиться только на характеристиках, специфичных для малых РНК и их мишеней [2].

Для отбора признаков был использован метод «Случайный лес» из библиотеки Scikit-learn. Этот метод основан на построении ансамбля решающих деревьев. Важным для данной работы является поле «feature_importance_» которую содержит обученный случайный лес. Значение данной переменной показывает важность каждого признака в предсказании целевой переменной, позволяя определить, какие из них оказывают наибольшее влияние на модель и являются наиболее информативными.

Результаты анализа показали, что наиболее значимыми признаками являются ACC, GCT и CCT, соответствующие различиям частот тринуклеотидов, а также признаки, связанные с гомологией и длиной последовательности. На основе работы «Случайного леса» было отобрано 45 наиболее важных признаков [3]. Анализ данных показал существенный дисбаланс между классами, что может негативно сказываться на работе моделей машинного обучения. Для уменьшения разницы в классах было решено случайным образом отобрать экземпляры нулевого класса, сократив дисбаланс до разницы в пять раз [3]. Для сокращения размерности пространства признаков был применен метод главных компонент (PCA). Основная цель метода состоит в том, чтобы проецировать исходные признаки на новое пространство меньшей размерности, при этом сохраняя максимально возможное количество дисперсии данных. Пространство признаков было сжато до 35-ти значений [3].

РАЗРАБОТКА АЛГОРИТМА

В рамках задачи машинного обучения решалась задача бинарной классификации. Для решения данной задачи была разработана нейронная сеть с 35 входными нейронами. Количество скрытых слоев и число нейронов на каждом скрытом слое были подобраны с использованием метода

автоматического подбора гиперпараметров. Каждая конфигурация модели была обучена дважды для исключения случайного влияния на хорошее предсказание. После процесса автоподбора была выбрана наилучшая архитектура: 288, 144, 96, 72, 57 и 48 нейронов на последовательно идущих скрытых слоях. В качестве функции активации на скрытых слоях использовалась функция ELU. На выходном слое два нейрона с функцией активации Softmax [3].

СНЯТИЕ МЕТРИК КАЧЕСТВА И СРАВНЕНИЕ С TARGETRNA3

Полученные метрики на тестовом наборе данных представлены в таблице.

Метрики качества Model и TargetRNA3

Алгоритм	AUC	F1	MCC
TargetRNA3	0,75	0,35	0,28
Model	0,90	0,55	0,48

Метрика AUC (Area Under the ROC Curve) представляет собой площадь под кривой ROC (Receiver Operating Characteristic). Чем выше значение AUC, тем лучше качество модели. ROC-AUC кривая строится путем построения графика зависимости True Positive Rate (TPR), вычисляемого по формуле (2), от False Positive Rate (FPR), вычисляемого по формуле (3), при различных порогах классификации.

F1-мера используется как метрика оценки моделей классификации. Она позволяет учитывать, как количество истинно положительных, так и ложноположительных результатов, и вычисляется по формуле (4).

MCC представляет собой коэффициент корреляции между предсказаниями модели и фактическими значениями, учитывая все четыре категории (TP, TN, FP, FN). MCC является более надежной метрикой, чем F1-мера в случае несбалансированных классов, что и наблюдается в данной работе, и вычисляется по формуле (5).

$$TPR = \frac{TP}{TP+FN}, \quad (2)$$

$$FNR = \frac{FP}{TN+FP}, \quad (3)$$

$$F1 = \frac{2*TP}{2TP+FP+FN}, \quad (4)$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}, \quad (5)$$

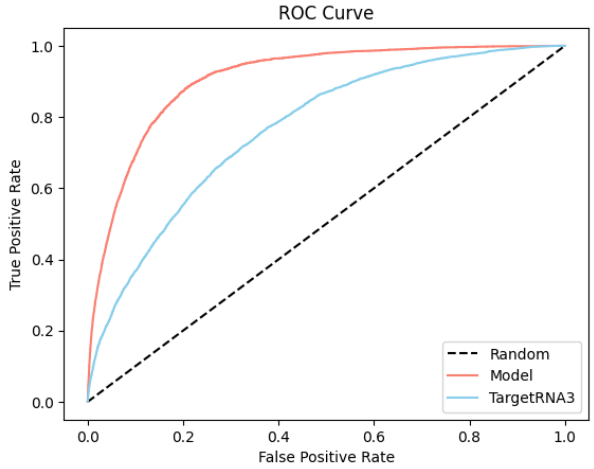
где TP – количество истинно положительных предсказаний,

TN – количество истинно отрицательных предсказаний,

FN – количество ложноотрицательных предсказаний,

FP – количество ложноположительных предсказаний.

Также данная модель была протестирована на всем наборе данных, и исходя из полученных предсказаний была построена ROC-AUC кривая. Значения TPR и FPR вычисляются согласно формулам (2) и (3). Model – алгоритм, представленный в данной работе. TargetRNA3 – вероятности предсказаний модели TargetRNA3. Метрики AUC на полном наборе данных составляют 0,906 и 0,767 для Model и TargetRNA3 соответственно, рисунок [3].



ROC-AUC кривая

ЗАКЛЮЧЕНИЕ

Несмотря на то, что в работе, посвященной TargetRNA3, одной из моделей-кандидатов была нейронная сеть, ее архитектура и структура остаются неясными. Однако данная работа демонстрирует, что при правильном подборе архитектуры нейронной сети можно достичь значительного улучшения метрик качества и качества классификации регуляторных мишеней прокариотической РНК.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. *Tjaden B.* TargetRNA3: predicting prokaryotic RNA regulatory targets with machine learning. // Harvard Dataverse. 2023. DOI: 10.7910/DVN/2Q8YRF
2. TargetRNA3: predicting prokaryotic RNA regulatory targets with machine learning [Электронный ресурс]. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-03117-2 (дата обращения: 17.04.2024).
3. RegulatoryRNAHunter [Электронный ресурс]. URL: https://github.com/InnerShadow/RegulatoryRNAHunter/tree/main (дата обращения: 17.04.2024).