О методах машинного обучения в задаче предсказания промоторов

Дюкова Анастасия Петровна anastasia.d.95@gmail.com
Москва, ФИЦ ИУ РАН

Дюкова Елена Всеволодовна edjukova@mail.ru
Москва, ФИЦ ИУ РАН

X МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ
«МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ И БИОИНФОРМАТИКА»
14-17 октября 2024 г., Пущино, Россия

Задача gene promoter prediction

Геномы – символьные последовательности в алфавите {A,C,G,T}, буквы которого соответствуют нуклеотидам ДНК. Одним из основных этапов расшифровки геномных последовательностей является нахождение в них особых структурных единиц – генов. Принципиально важным является нахождение границ генов в последовательности генома и в связи с этим определение специальных участков генома (регуляторных участков экспрессии гена), называемых промотерами. Существующие походы решения задачи распознавания промотера *(*англ.: gene promoter prediction*)* недостаточно эффективны и интерпретируемы.

Представляется перспективным применение методов логического анализа и классификации данных для рассматриваемой задачи. Нахождение адекватных способов поиска промотеров важно ещё и потому, что позволит предложить в перспективе новые способы идентификации других структурных сегментов геномов.

Подходы к решению задачи gene promoter prediction

Для решения задачи gene promoter prediction используются методы машинного и глубокого обучения.

По сути решается хорошо известная в интеллектуальном анализе данных задача классификации по прецедентам. Под прецедентной (обучающей) информацией понимается совокупность примеров изучаемых участков генома, в которой каждый пример представлен в виде числовой последовательности, полученной на основе наблюдения и измерения ряда его характеристик. Такие характеристики называются признаками. В случае бинарной классификации примеры делятся на два класса (класс положительных и класс отрицательных примеров). К положительным примерам относятся промотеры изучаемого организма, а к отрицательным — другие участки генома этого организма, например, экзоны или интроны. Требуется на основе анализа обучающей выборки уметь отличать промотерный участок геномной последовательности организма от непромотерного участка этого организма.

Понятие модельного организма. Методика формирования признаков

Модельный организм в генетике — это организм, который изучают для понимания основных генетических процессов благодаря его доступности, и простоте генома. В большинстве работ по распознаванию промотеров в качестве модельных организмов рассматриваются кишечная палочка, человек, дрозофила и арабидопсис. В настоящей работе вопросы применения методов машинного обучения в задаче gene promoter prediction изучались на примере Дрозофилы фруктовой (Drosophila melanogaster).

Для формирования признаковых описаний изучаемых областей генома широко используется показатели, именуемые *k*-мерами и представляющие собой всевозможные упорядоченные наборы длины *k* из символов A, C, G и T (слова длины k в 4-х буквенном алфавите). Каждая k-мера соответствует некоторому признаку и частота встречаемости k-меры в рассматриваемой области генома задаёт значение соответствующего признака. В результате данные преобразуются в последовательности целых чисел.

Результаты предыдущих исследований для Drosophila melanogaster

Для рассматриваемой задачи применяются традиционные алгоритмы машинного обучения, такие как метод опорных векторов (SVM), логистическая регрессия, случайный лес, дерево решений, градиентный бустинг и многослойный перцептрон, а также модели классификаторов, основанные на глубоком обучении, например, свёрточная нейронная сеть.

Большинство исследований проводится на сбалансированных выборках (СВ), т.е. на выборках с одинаковым числом положительных и отрицательных примеров. Для оценки качества классификации наиболее часто используются t-fold кросс-валидация (CV), в частности, leave-one-out CV тест. На СВ наилучшая точность классификации 97,5 % получена методом SVM и применением нейронной сети соответственно в работах:

1. Zhang, M., Jia, C., Li, F., Li, C., Zhu, Y., Akutsu, T., Webb, G. I., Zou, Q., Coin, L. J. M., Song, J. Briefings in bioinformatics., 2022. Vol. 23(2), bbab551.

2. Zhu, Y., Li, F., Xiang, D., Akutsu, T., Song, J., Jia, C. Briefings in bioinformatics., 2021. Vol. 22(4), bbaa299.

Основная цель работы

Основная цель работы – исследование возможности применения логических методов анализа и классификации данных в задаче gene promoter prediction на примере организма Дрозофилы фруктовой (Drosophila melanogaster). Логический подход базируется на поиске информативных фрагментов в признаковых описаниях прецедентов и ориентирован на обработку целочисленной информации низкой значности. Искомые фрагменты, называемые элементарными классификаторами, хорошо интерпретируемы и позволяют отличать промотеры от других областей генома, однако их поиск требует больших временных затрат.

Рассматривались вопросы эффективности применения логического подхода к рассматриваемой задаче в двух следующих случаях**:**

- с использованием традиционной методики формирования признаков на основе *k* -мер;
- прямое применение классификатора к исходным символьным данным (без формирования признаков на основе *k* -мер).

Результаты проведенного исследования (1)

Для формирования выборки, содержащей 16000 примеров промотерных последовательностей организма Drosophila melanogaster и 42000 отрицательных примеров, представляющих собой участки экзонов, использовались соответственно базы данных EPDNnew (https://epd.expasy.org/epd) и Flybase (https://flybase.org). Длина промотерной последовательности составила 300 пар нуклеотидов (п.н.). Выбранная длина считается оптимальной.

Первоначально для формирования признаков использовались *k*-меры. Тестировались следующие алгоритмы машинного обучения:

- случайный лес (RF);
- логистическая регрессия (LR);
- градиентный бустинг (LightGB, XGBoost, CatBoost);
- логический классификатор (REC), основанный на поиске часто встречающихся фрагментов описаний прецедентов, позволяющих различать прецеденты из разных классов.

Результаты проведенного исследования (2)

Качество классификации оценивалось двумя функционалами: сбалансированной точностью (Q1) и ROC-AUC (Q2). Применялась методика многократного случайного разбиения исходной выборки на обучающую и тестовую подвыборки в пропорции 4:1 (с усреднением результатов запусков). В экспериментах использовалась авторская реализация логического классификатора. Реализации других тестируемых алгоритмов брались из известной библиотеки машинного обучения scikit-learn и запускались с базовыми параметрами.

Результаты счёта классических моделей с применением методики формирования признаков на основе *k*-мер приведены в таблице 1, в которой для каждого рассмотренного случая указано максимальное значение параметра *k* , обозначаемого *maxk*. Для всех классификаторов наилучшие результаты получены при *max k* = 4, т.е. при *k* равном 1, 2, 3 и 4. В этом случае для описания прецедентов использовались 340 признаков. Наивысшую точность классификации показал алгоритм CatBoost (Q1 = 87,6%, Q2 = 94,8%). На втором месте алгоритм XGBoost (Q1 = 86,7%, Q2 = 94,2%). Другие тестируемые классификаторы показали сравнимые между собой результаты, уступив показателям CatBoost и XGBoost.

Табл. 1. Точность классификации Q1/ Q2 на исходной выборке (16000, 42000)

maxk **(число признаков)**	**LR**	**RF**	**LightGB**	**Catboost**	**XGBoost**
2 (20)	0,762/0,884	0,789/0,908	0,778/0,892	**0,792/0,903**	0,789/0,898
3 (84)	0,805/0,914	0,804/0,926	0,807/0,918	**0,833/0,933**	0,826/0,927
4 (340)	0.854/0.933	0.844/0.933	0.846/0.929	**0.876/0.948**	0.867/0.942
5 (1364)	0.839/0.937	0.794/0.928	0.822/0.928	**0.851/0.946**	0.843/0.940

Результаты проведенного исследования (3)

В таблице 2 приведены результаты тестирования логического классификатора REC при *maxk* = 4. Из-за временных ограничений применялся ансамблевый подход на основе бэггинга над экзонами. Кроме того проводилась предварительная обработка данных с целью «корректного» снижения значности исходных целочисленных значений признаков. При корректной перекодировке данных описания обучающих объектов из разных классов остаются различимыми. Для уменьшения времени счёта первоначальное число признаков было сокращено до 150 за счёт выделения наиболее информативных признаков. Классификатор REC уступил CatBoost по обоим показателям Q1 и Q2 примерно на 4%.

Результаты проведенного исследования (4)

Выделение информативных признаков (информативных мер) проводилось двумя способами. Первый способ, названный методом CVP, основан на анализе информативности отдельных значений признаков (Дюкова Е. В., Песков Н. В. 2002 г.), а второй – на методе, встроенном в CatBoost. В пересечении выделенных двумя указанными способами мер оказалось 74 меры. В первом случае в первую десятку по убыванию информативности вошли символьные наборы TT, T, TTT, TTA, GGA, TA, TTTT, C, ATT, ATTT, а во втором – наборы TT, GGA, AGGA, T, AAA, TA, CCT, CC, AGAG, ATG. Нетрудно видеть, что в обоих случаях были выделены меры TT, T, GGA, TA. Следует отметить, что 4-меры TTTT и ATTT, выделенные методом CVP, были также отмечены как наиболее информативные и другими авторами.

Табл. 2. Сравнительный анализ классификаторов REC и CatBoost на исходной выборке (16000, 42000) при $maxk = 4$

	REC	CatBoost
	С понижением значности признаков	Без понижения значности признаков
без отбора информативных признаков	_	**0,876 / 0,948**
с отбором информативных признаков по методу CVP	0,829 / 0,904	**0,866 / 0,940**
с отбором информативных признаков по методу CatBoost	0,831 / 0,905	**0,875 / 0,947**

Результаты проведенного исследования (5)

Дополнительно было проведено тестирование классификаторов REC, CatBoost и XGBoost на исходных символьных описаниях промотеров и экзонов с длиной 300 п.н. (без формирования признаковых описаний объектов на основе *k*-мер). Положительные и отрицательные примеры объектов, содержащие в совокупности 3000 прецедентов, отобранных случайным образом, были разбиты на участки, состоящие из 20 п.н. К получившимся подвыборкам напрямую применялись указанные алгоритмы. Результаты счёта каждого классификатора усреднялись по 10 запускам. Итоговая точность для REC составила Q1=0.831, Q2=0.943. Время счёта – 15 минут. Точность XGBoost также оказалась выше по ROC-AUC, но немного понизилась по показателю Q1 и составила Q1=0.854, Q2=0.948. Для CatBoost получены следующие результаты: Q1=0.863, Q2=0.951.

Таким образом, при применении напрямую к исходным символьным описаниям промотеров и экзонов точность классификатора REC повысилась существенно, а классификаторы XGBoost и CatBoost показали незначительные изменения по функционалам Q1 и Q2.

Заключение

Впервые для решения задачи gene promoter prediction применены методы логического анализа и классификации данных. Приведены результаты экспериментов на несбалансированной выборке большого объёма, при этом рассмотрен как традиционный способ формирования признаков, использующий *k*-меры, так и методика прямого применения классификатора к исходным данным. Показано, что во втором случае качество логической классификации существенно выше и составляет 94,3% по ROC-AUC с использованием ансамблевого подхода. Наилучший результат, а именно, точность по ROC-AUC равную 95,1% , показал классификатор Catboost при прямом применении к исходной выборке. При традиционном способе формирования признаков точность Catboost равна 94,8%.

Слайд 14