ПОДХОД К АВТОМАТИЗИРОВАННОМУ ВЫБОРУ НАИЛУЧШЕЙ МОДЕЛИ В ЗАДАЧЕ ПОЛНОГЕНОМНОГО ПОИСКА АССОЦИАЦИЙ

Р. С. Сергеев, В. В. Сатаневский

Объединенный институт проблем информатики НАН Беларуси
Минск, Беларусь
e-mail: roma.sergeev@gmail.com

В работе рассматривается подход к автоматизированному выбору наилучшей модели в задаче поиска ассоциаций генотип/фенотип, включая методологию оценки параметров и отбора значимых признаков. Особое внимание уделяется вопросам оценки качества результатов и проблеме переобучения.

Ключевые слова: полногеномный поиск ассоциаций; биоинформатика; лекарственная устойчивость; скользящий контроль.

APPROACH TO AUTOMATED SELECTION OF SIGNIFICANT FEATURES IN GENOME-WIDE ASSOCIATION STUDIES

R. S. Sergeev, U. V. Sataneuski

United Institute of Informatics Problems of NAS Belarus
Minsk, Belarus

This paper is devoted to the approach to automated selection of best performing model in genome-wide association studies, estimation of the model parameters and using it for selection of significant features. Special attention is payed to fair evaluation of prediction quality and prevention of overfitting.

Keywords: genome-wide association study; bioinformatics; drug resistance; cross-validation.

ВВЕДЕНИЕ

Стремительное развитие методов секвенирования нового поколения породило рост геномных исследований. В этом контексте важное место занимают задачи геномного поиска ассоциаций, целью которых является изучение связи генотипа вых организмов с различными фенотипическими проявлениями. В подобных исследованиях сравниваются геномы организмов, являющихся носителями некоторого знака, и геномы организмов, где этот признак никак не проявляется либо эти ния носят иной характер. Современные методы машинного обучения позволяют матизировать этот процесс, однако существует ряд особенностей, которые затрудняют выявление достоверных закономерностей. В докладе будет рассмотрена задача поиска геномных маркеров лекарственной устойчивости микобактерий туберкулеза, а также

некоторые подходы к ее решению. Исходные данные для исследования были предоставлены РНПЦ пульмонологии и фтизиатрии Министерства здравоохранения лики Беларусь.

ПОСТАНОВКА ЗАДАЧИ

Рассматривается задача полногеномного поиска ассоциаций, где анализируются однонуклеотидные полиморфизмы в последовательностях ДНК микробактерий туберкулеза. Цель исследования заключается в нахождении таких участков генома, мутации в которых влияют на наличие либо отсутствие лекарственной устойчивости к определенному препарату. Для идентификации мутаций использовалось сравнение геномных последовательностей с референсной последовательностью H37Rv.

Подходы к решению. В решении подобных задач часто используют методы шинного обучения, позволяющие отбирать значимые признаки и строить модели для предсказания лекарственной устойчивости новых образцов [1]. Одна из отличительных черт рассматриваемой задачи заключается в наличии малого числа секвенированных геномов и большого числа анализируемых геномных вариаций. Как известно, в таких ситуациях велик риск переобучения. В связи с этим, важным этапом решения задачи становится разработка протокола тестирования, позволяющего корректно сравнивать результаты разных методов. Также хотелось, чтобы подходы к решению задачи проверялись автоматически, а результаты тестирования не зависели от конкретных гиперпараметров рассматриваемых моделей.

Более конкретно требования к тестирующему окружению можно сформулировать следующим образом:

а) Независимость от гиперпараметров. В вычислительных экспериментах задается лишь структура модели, но не указываются конкретные значения параметров и гиперпараметров. Это позволяет сравнивать структуру моделей, а не умение настраивать параметры.

б) Подбор параметров эксперимента проводится независимо от тестирования. Это означает, что тестовые данные не участвуют при поиске оптимальных значений гиперпараметров, т. е. тестирование проводится ровно один раз, после того как все параметры перебраны. В противном случае мы будем иметь информацию о качестве модели на тестовых данных, что увеличит риск переобучения. Обычно этому уделяют не так много внимания, поскольку на выборках с большим числом объектов рост метрик качества модели (точности, полноты, *F*-меры, правильности) с большей вероятностью свидетельствуют об улучшении модели, а не о переобучении.

в) Устойчивость к разбиению. Оценка качества должна слабо зависеть от разбиения данных на обучающую и тестовую выборки.

Структура тестирующего окружения, сформированная в соответствии с приведенными требованиями, представлена на рис. 1.

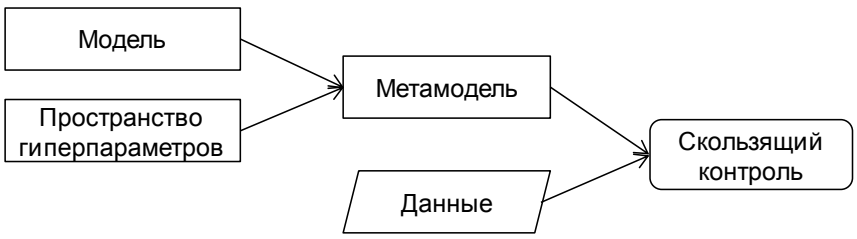

Рис. 1. Структура тестирующего окружения

1) На вход поступают данные по некоторому лекарству, модель, а также пространство неинициализированных параметров и гиперпараметров этой модели.

2) Формируется метамодель, у которой отсутствуют собственные гиперпараметры, а переданные на предыдущем шаге модель и пространство возможных значений гиперпараметров рассматриваются в качестве внутренних параметров самой метамодели.

3) Метамодель обучается и тестируется посредством скользящего контроля. На выход поступают результаты тестирования – метрики качества метамодели и отобранные ею признаки.

Для обучения и тестирования метамодели применяется двухэтапная процедура скользящего контроля по пяти блокам. На внешнем цикле скользящего контроля оценивается обобщающая способность метамодели. При этом тестовый блок каждого разбиения используется для выполнения предсказаний и вычисления метрик качества, а обучающие блоки образуют выборку, на которой выполняется вложенный цикл скользящего контроля для подбора оптимальных параметров метамодели. На этой же выборке происходит обучение внутренней модели с использованием найденных оптимальных значений параметров метамодели (гиперпараметров внутренней модели).

Заметим, что такой способ тестирования удовлетворяет заявленным требованиям: метамодель сама является моделью, не имеющей гиперпараметров, что влечет выполнение условия а), благодаря использованию метода скользящего контроля удовлетворяются условия б), -в).

Для большинства алгоритмов машинного обучения от правильности задания перпараметров зависит качество результирующей модели. Чтобы оценить, какой метод машинного обучения работает лучше, значения гиперпараметров выбираются достаточно близкими к оптимальным (в той мере, в которой это возможно), после чего алгоритмы, инициализированные наилучшими найденными значениями метров, сравниваются между собой. В рассматриваемой задаче оценка оптимальных значений гиперпараметров выполнялась с использованием процедуры байесовской оптимизации [2]. В отличие от более классических методов подбора гиперпараметров, таких как поиск по сетке, это позволяет осуществлять выбор вариантов новых ний гиперпараметров, опираясь на результаты предыдущих проверок, что обеспечивает сходимость за небольшое число итераций.

РЕАЛИЗАЦИЯ

Программная реализация основных процедур выполнена на языках Python и C с привлечением сторонних библиотек: hyperopt (оптимизация гиперпараметров), scikit-learn (реализация основных методов машинного обучения и вычисления метрик качества), xgboost (методы градиентного бустинга) и scipy (проверка статистических гипотез при определении значимости результатов).

ПРОВЕДЕННЫЕ ЭКСПЕРИМЕНТЫ И ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ

Чтобы при выполнении вычислительных экспериментов не приходилось каждый раз задавать всевозможные варианты параметров, относительно несложные пары (модель и пространство параметров) были объединены в примитивы, которые послужили составными блоками для формирования итоговых моделей. Опишем примитивы, которые были использованы в моделях, получивших наилучшие оценки качества.

- Примитив «Выбор». Используется в случае, когда имеется несколько моделей, но неизвестно, какая из них лучше. Пространством параметров примитива являются варианты различных моделей.
- Примитив «Логистическая регрессия». В качестве модели используется регуляризованная логистическая регрессия. Перебираемые параметры: коэффициент регуляризации $\lambda \in \left[e^{-15}, e^{15}\right]$, метод регуляризации $\{l_1, l_2\}$, веса классов.
- Примитив «Случайный лес». В качестве модели выступает случайный лес из 100 деревьев.
- Примитив «Градиентный бустинг». В качестве модели используется фикатор XGBClassifier библиотеки xgboost [3]. Перебираемые параметры: количество деревьев (от 1 до 200), максимальная глубина дерева (от 1 до 13), минимальное число листовых объектов (от 1 до 6), доля набора данных, используемая при обучении одного дерева (от 0,5 до 1), минимальное изменение функции потерь, необходимое для разделения листового узла (от 0,5 до 1).
- Примитив «Базовая модель». Использует примитив «Выбор», параметрами которого выступают примитивы «Логистическая регрессия», «Случайный лес» и «Градиентный бустинг».
- Примитив «Статистический метод отбора признаков». Примитив служит для выбора одного из двух методов, основанных на проверке статистических гипотез: хи-квадрат и ANOVA. Перебираемые параметры: статистический критерий (хи-квадрат или ANOVA), количество результирующих признаков $k \in [e^{-15}, e^{15}]$.
- Примитив «Отбор признаков на основе модели». В качестве модели выступает метод, отбирающий признаки на основе важности, вычисляемой с помощью случайного леса или логистической регрессии. Перебираемые параметры: модель, вычисляющая порог, и декорируемая модель, использующая отобранные признаки.
- Примитив «Отбор признаков». Выбор между примитивами «Отбор признаков на основе модели» и «Статистический метод отбора признаков».
- Примитив «Сеть релевантности признаков». В качестве модели используется сеть релевантности признаков [4]. Перебираемые параметры: коэффициенты $\lambda \in \left[e^{-15}, e^{15}\right]$ и $\alpha \in (0, 1)$.

Использование примитивов оказалось очень удобным: например, если в итоговой модели необходимо было использовать один из нескольких базовых алгоритмов машинного обучения (однако на этапе определения модели не было известно, какой из них сработает лучше), применялся примитив «Базовая модель» вместо указания конкретных моделей и их допустимых параметров.

В таблице представлены результаты некоторых экспериментов с моделями, построенными на основе комбинаций примитивов:

A) «Базовая модель».

B) «Отбор признаков» + «Базовая модель».

C) «Сеть релевантности признаков» + «Базовая модель».

Результаты экспериментов с моделями на основе комбинаций примитивов

Препарат	Эксперимент	Число истинно положительных	Число истинно отрицательных	Число ложноположительных	Число ложноотрицательных	*F*-мера	Правильность
AMIK	A	54	67	4	11	0,88	0,89
	B	53	68	3	12	0,88	0,89
	C	53	67	4	12	0,87	0,88
CAPR	A	55	54	5	17	0,83	0,83
	B	55	53	6	17	0,83	0,82
	C	56	51	8	16	0,82	0,82
OFLO	A	59	54	6	17	0,84	0,83
	B	63	55	5	13	0,88	0,87
	C	61	54	6	15	0,85	0,85

Как видно из таблицы, полученные результаты оказались достаточно похожи. При этом наилучшая прогностическая способность была продемонстрирована комбинацией примитивов «Отбор признаков» и «Базовая модель» (эксперимент B). По итогам экспериментов с остальными примитивами можно заключить, что применение базовых методов отбора признаков явилось более предпочтительным в рассматриваемой задаче.

Проиллюстрируем также значения доли верных предсказаний, которые получались в процессе поиска оптимальных параметров метамоделей на внутреннем и внешнем скользящем контроле (рис. 2).

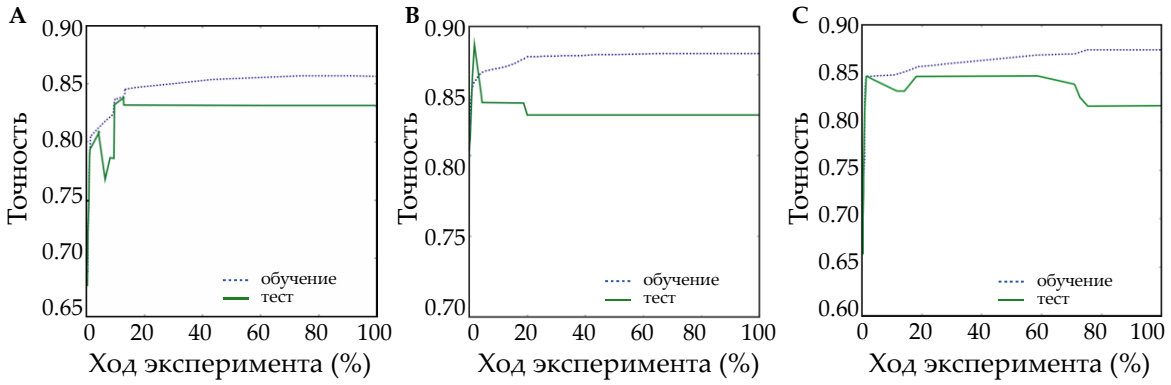

Рис. 2. Графики, иллюстрирующие сходимость процесса подбора параметров в обучающем и тестовом циклах скользящего контроля для препарата CAPR

По приведенным графикам видно, что качество при подборе параметров посредством скользящего контроля на обучающей выборке заметно выше, чем итоговое качество при тестировании. В свою очередь, это свидетельствует об обоснованности построения двухуровневой процедуры тестирования для получения объективных результатов сравнения алгоритмов.

ЗАКЛЮЧЕНИЕ

В работе было проведено сравнение различных методов машинного обучения и их комбинаций на примере задачи полногеномного поиска маркеров лекарственной устойчивости микобактерий туберкулеза к противомикробным препаратам второго ряда. Разработано программное обеспечение для комбинирования методов и автома-

тического выбора наилучшей модели, предложена методология объективной оценки и сравнения результатов.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Алгоритмы поиска мутаций лекарственной устойчивости в геномах микобактерий туберкулеза / Р. С. Сергеев [и др.] // Информатика. 2016. № 1(49). С. 75–91.
2. Hyperopt: a Python library for model selection and hyperparameter optimization / J. Bergstra [et al.] // Comput. Sci. Disc. 2015. Vol. 8. № 1. doi: 10.1088/1749-4699/8/1/014008.
3. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. P. 785–794.
4. High-Dimensional Structured Feature Screening Using Binary Markov Random Fields / J. Liu [et al.] // JMLR workshop and conference proceedings, 2012. Vol. 22. P. 712–721.