Методы машинного обучения для предсказания распределения генетической информации по популяциям

Попов Дмитрий Олегович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: o3vh7etsbccg@maskedmails.com

Научный руководитель *— Дьяконов Александр Геннадьевич*

Оценка происхождения людей по их генетической информации является важной задачей для коррекции результатов биологических исследований. Регулярно проводятся исследования, результаты которых оказываются невоспроизводимы на других континентах и в других популяциях из-за неизвестной заранее связи исследуемого явления с генотипом человека. Именно поэтому проведение тестирования на далёких друг от друга популяциях является важным критерием для определения уровня доверия к результатам экспериментов. Надёжными способами оценивать происхождение людей являются математические методы обработки генетической информации в виде значений нуклеотидов в заранее выбранных участках генома — позициях однонуклеотидных полиморфизмов.

В биологических исследованиях наиболее часто используются алгоритмы кластеризации с обучением без учителя, основанные на EM-алгоритме. Примерами самых популярных служат ADMIXTURE [1–2], EIGENSTRAT [3] и sNMF [4]. Используемые алгоритмы имеют одинаковую основополагающую идею, а различаются эвристиками для ускорения процесса оптимизации. Будучи методами обучения без учителя, они не способны принимать во внимание размеченные выборки данных, например, собранных в рамках фундаментального исследования «1000 Геномов» [5]. Кроме того, в сравнении с результатами данной работы, они обладают низкой скоростью из-за невозможности прямой оптимизации, рассчитаны на малое (менее 10) количество популяций и малое (до 10^4) количество признаков. Для возможности включения в метод решения данных о нескольких десятках популяций, представленных 10^6 признаками на объект необходима разработка более эффективных методов.

В рамках данной работы были представлены локальный и глобальный подход к обработке данных, а также применен ряд метрических, линейных, вероятностных и ансамблевых методов обучения с учителем. Кроме того, был представлен авторский способ тести-

рования методов, учитывающий специфику задачи и основанный на знаниях из предметной области, который таким образом лучше отражает качество решения на реальных данных.

Для сравнения методов использовался публичный набор данных проекта «1000 Геномов», объекты которого представляют собой последовательности значений пар нуклеотидов людей, вошедших в исследование, отнесённых каждый к одной из 25 популяций. Способ тестирования заключается в применении алгоритмов к смешанным геномам, искусственно синтезированным из объектов отложенной выборки согласно теоретическим представлениям наследования информации в генетике.

Проведённые эксперименты включали в себя сравнительное тестирование рассматриваемых в работе алгоритмов на полной выборке данных с использованием предложенного метода тестирования. Также они сравнивались с приведёнными выше уже используемыми в биологических исследованиях методах обучения без учителя на выборке с усечёнными множествами признаков, поскольку на полной готовые методы работали неразумно долгое время. Наибольшее качество показал ансамблевый метод, включающий в себя локальный и глобальный подходы. На меньшей выборке он продемонстрировал более быстрое и качественное решение, чем эталонные алгоритмы.

Литература

1. Alexander D. H., Lange K. Enhancements to the admixture algorithm for individual ancestry estimation. BMC Bioinformatics 2011 12:246.
2. Alexander D., Novembre J., Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research 2009, 19:1655-1664.
3. Price A. L., Patterson N. J., Plenge R. M., Weinblatt M. E., Shadick N. A., Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.
4. Eric F., François M., Théo T., Guillaume B., Olivier F.: Fast and Efficient Estimation of Individual Ancestry Coefficients. Genetics April 1, 2014 vol. 196 no. 4 973-983.
5. Официальная страница проекта «1000 Геномов»: `https://www.internationalgenome.org/`