

УДК 004.942

ПРИМЕНЕНИЕ БАЙЕСОВСКОЙ ОПТИМИЗАЦИИ ДЛЯ ВЫВОДА ДЕМОГРАФИЧЕСКОЙ ИСТОРИИ ПОПУЛЯЦИЙ ПО ГЕНОМНЫМ ДАННЫМ

Хужин П.А. (Университет ИТМО)

Научные руководители – к.т.н., доцент ФИТиП Ульяновцев В.И.,

Носкова Е.Э.

(Университет ИТМО)

Построение демографической модели популяций является одной из важных проблем популяционной генетики. Настоящая работа посвящена разработке алгоритма, основанного на байесовской оптимизации и его применению для вывода демографических историй для четырех и пяти популяций по геномным данным.

Введение. Демографическая история (модель) — это детальная история развития популяций, включающая такие параметры, как численность популяций, время разделения и темпы миграций. Существует несколько решений для предсказания демографических историй по геномным данным. Одним из наиболее перспективных методов является аппроксимация моментов случайного процесса. Этот метод был реализован в программном обеспечении *moments* и поддерживает вывод историй до пяти популяций. Исследователь сам задает сценарий и затем находит параметры, дающие максимальное значение правдоподобия с реальными генетическими данными.

Данная задача оптимизации была решена различными методами, например, в *moments* были использованы различные алгоритмы локального поиска. Недавно был представлен первый алгоритм глобальной оптимизации, основанный на генетическом алгоритме — *GADMA*. Однако вычисление значения правдоподобия имеет экспоненциальную сложность от числа популяций, и задача эффективного и быстрого поиска параметров демографической истории для четырех и особенно для пяти популяций до сих пор является вычислительно-сложной. Именно по этой причине представленный метод генетического алгоритма поддерживает только до трех популяций.

Основная часть. Одним из методов для поиска оптимума у сложновычислимых функций является байесовская оптимизация. Это один из лучших подходов для оптимизации в непрерывном пространстве размерности меньше двадцати. В данной работе мы применили алгоритм байесовской оптимизации к задаче поиска параметров демографической истории для четырех и пяти популяций в *moments*. Мы провели подбор гиперпараметров метода, таких как ядро, число исходных точек и учитывание шума, исходя из проведенных экспериментальных исследованиях и знаний о предметной области, полученных во время разработки *GADMA*. Были проведены первичные экспериментальные исследования для выявления эффективности полученного метода.

Выводы. Метод является первым эффективным алгоритмом глобального поиска параметров демографической истории для четырех и пяти популяций из геномных данных. Байесовская оптимизация показала себя компромиссом между скоростью поиска и точностью полученного решения.

Хужин П.А. (автор)

Подпись

Ульянцев В.И. (научный руководитель)

Подпись

Носкова Е.Э. (научный руководитель)

Подпись