Отчет о решении задачи по биоинформатике
Панкратов Виктор, M05-104a

1 Постановка задачи

Задано множество конечных последовательностей $\{t_i\}_{i=1}^{n}$,n - не фиксированное натуральное число, состоящих из четырех типов символов:

$$\forall i \in \{1 \dots n\}\ t_i \in \{A, C, G, T\}$$

Требуется проверить возможность определения конца последовательности - построения некоторой функции $\{x\}_1^k \to y \in \{0, 1\}$, где x – некоторая подпоследовательность $\{t\}_1^n$ фиксированной длины, а y – индикатор близости к концу последовательности. В данной работе полагалось $x = \{t_i\}_{i=n-k}^{n} \to y(x) = 1$ и $x = t_{i\,i=n-k-K}^{n-K} \to y(x) = 0$ при достаточно больших K, определенных далее.

2 Базовые предположения

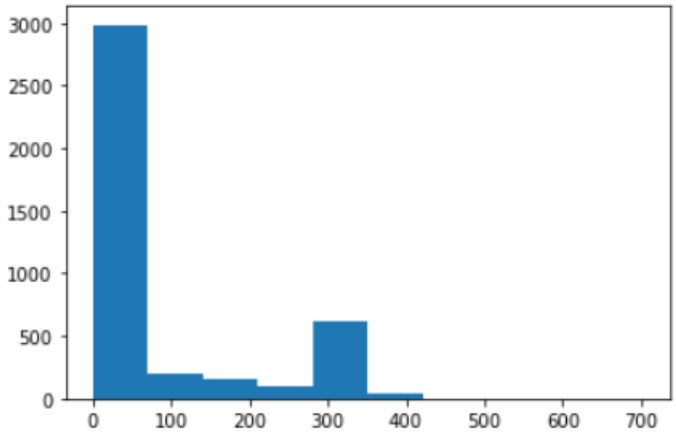
- Искомая функция принимает на вход K-элементную подпоследовательность с некоторым не зависящим от самой последовательности K
- Искомая функция не зависит от последовательности и зависит только от поданных на вход K символов. В частности, при определении терминаторов не необходимо учитывать начало последовательности.
- Предыдущее предположение также порождает следующее: если для некоторой подпоследовательности индикатор близости равен 1, то она не встречается в середине другой последовательности
- При невозможности достижения идеального результата ошибки первого и второго рода считаются равными по значимости, а задача переходит в задачу максимизации корректных предсказаний функции
- Функция должна быть детерминированной. Предположение имеет смысл только в условиях предыдущего

3 Данные

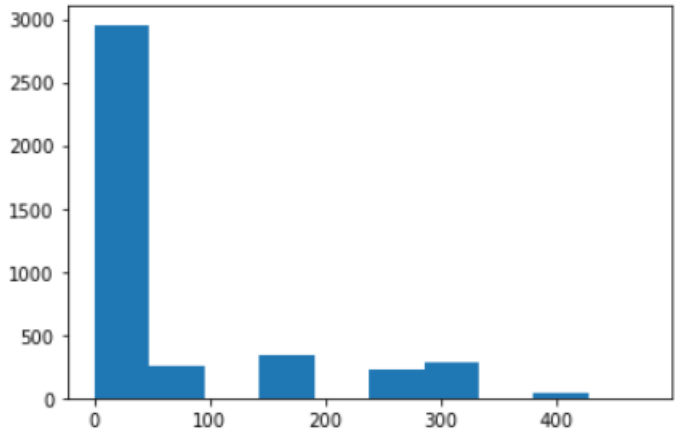
4116 последовательностей $\{t_i\}$ описанного выше вида. Пары (x, y) были составлены по указанному выше принципу для тех x, для которых по такому принципу было возможно определить y.

4 Решение

При получении задачи дополнительно предполагалось $K \leq 300$. Покажем, что в этом случае точно решить задачу нельзя: рассмотрим последние 300 символов каждой последовательности из данных. Попробуем найти их в виде подпоследовательности среди не конечных элементов других последовательностей. Результаты представлены на гистограмме ниже.



Почти такой же по характеру результат получается при увеличении K до 500.



Это вынуждает скорректировать задачу. Можно было рассматривать её только для последовательностей, концы которых не были найдены в других, но из их малого числа возникает предположение о том, что для таких последовательностей они могли быть найдены в других, не представленных в выборке. Теперь задача заключается именно в достижении минимальной ошибки.

4.1 Нейросетевой подход

Была построена простая нейросеть из нескольких слоев и обучена на парах (x, y). Эксперимент был повторен 5 раз. Тестовый набор содержал одинаковое число y равных 0 и 1. Разделение выборки на обучение, валидацию и тест делалось случайно при каждом запуске. Было получено, что в каждом эксперименте модель правильно предсказывала конец в $73 - 75$ процентах случаев.

4.2 Решающие деревья

Был построен классификатор на основе решающего дерева и обучен на парах (x, y). Тесты не повторялись, поскольку были проведены на всей выборке. Обученный на описанной выше ее части классификатор достиг 95.6 процентов истинных предсказаний на данных из середины последовательностей и 97.2 на данных из конца.

5 Вывод

Мы получили, что точно решить данную задачу невозможно, но тем не менее рассматриваемые подходы дали приближенное решение. Более того, окончательно примененный подход(деревья) позволил получить этот же результат в интерпретируемом виде. К большему качеству в работе я не стремился, поскольку примерно 1.5 процента ошибки получаются уже из-за совпадений подпоследовательностей с концами строк.

6 Ссылки

- Репозиторий с кодом
- Статья, из результатов которой получены используемые данные