

Скрытые Марковские модели переменного порядка для анализа данных ChIP-seq

Атаманова Анна, кафедра системного программирования СПбГУ, anne.atamanova@gmail.com

1 апреля 2015 г.

Аннотация

Здесь нужно кратко описать суть работы и результаты.

1 Введение

Дезоксирибонуклеиновая кислота (ДНК) является своеобразным кодом жизни. Эта молекула хранит и передает генетическую программу развития и функционирования живого организма. В то же время, все функции ДНК зависят от ее соединений с белками. Поэтому, изучение ДНК-белковых взаимодействий актуально и привлекательно. Chip-seq (chromatin immunoprecipitation - sequencing)[1] является одним из современных методов, позволяющим выделить участки ДНК связанные с конкретным белком (одинаково применим к разным белкам). Однако, по понятным причинам (сложный биологический эксперимент), погрешность данного метода не может быть нулевой, и безрассудная вера ему лишена смысла. По этому, обычно, к результатам подобных методов накладывается вероятностная модель. Конечно, это добавляет ряд существенных ограничений. Однако, в качестве неоспоримого плюса можно привести тот факт, что хорошо подобранная модель позволяет понять природу данных и изучить их свойства.

Итого, нашей задачей является нахождение модели по последовательности чисел выдаваемых Chip-seq.

В настоящее время, в качестве семейства искомых моделей, активное применение находит НММ (Hidden Markov Model)[2] второго порядка с Пуассоновским испусканием. Данное семейство допускает предположение о том, что каждое состояние (наличие/отсутствие белка в заданной части генома) зависит только от одного предыдущего. Можно ограничиться и более лояльным допущением о том, что состояние зависит от n предыдущих состояний, однако такое допущение резко увеличивает сложность модели ($O(2^n)$ параметров). Также, сложность заключается в подборе этого n и переобучении в случае, если не все состояния имеют одинаковые длины контекстов зависимости. Последнее замечание подводит к идее использования VОНММ (Variable Order Hidden Markov Model)[3]

2 Скрытые марковские модели переменного порядка

Марковский процесс порядка m - это случайный процесс, эволюция которого в каждый момент времени зависит только от $m - 1$ предыдущих состояний.

Другими словами, $X = (x_1, x_1, \dots, x_T)$ является марковской цепью порядка m , если $p(x_t | x_1^{t-1}) = p(x_t | x_{t-m+1}^{t-1})$

Таким образом, каждое следующее состояние определяется контекстом длины $m - 1$ и вероятностями перехода из него.

Скрытая марковская модель предполагает, что, помимо всего прочего, каждое состояние задает свою вероятностную модель (например, гауссиан), а в качестве наблюдений поступают не сами состояния, а лишь испускания из них y_1, \dots, y_T (1- пример НММ порядка 2).

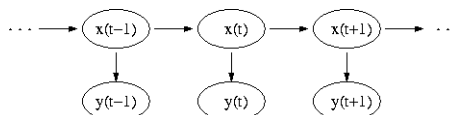


Рис. 1: НММ order 2

Скрытая марковская модель переменного порядка - обобщение НММ разрешающее иметь контекстам разные длины.

Итого, скрытая марковская модель переменного порядка определяется контекстами, вероятностными распределениями переходов и вероятностными распределениями испусканий для каждого состояния.

Перейдем к обучению модели по цепочке наблюдений.

Обычно обучение $n-hmm$ m -го порядка сводится к n^m-hmm , которая обучается алгоритмом Баума-Велха (частный случай ЕМ алгоритма, который итеративно максимизирует правдоподобие модели).

Для хранения контекстов используется дерево (бор), в котором вершины являются строками, ребра буквами. Корень - это пустой контекст, а каждый ребенок вершины является уточнением контекста родителя на состояние ребра. Итого, каждая внутренняя вершина имеет ровно n потомков. Листья - главные контексты.

Заметим, что, если дети какой-то вершины имеют одинаковые вероятности переходов, то они контекст вершины не уточняют, и их можно обрезать.

Так же заметим, что достаточно хранить только листья и распределение переходов на них, и, в случае необходимости, пересчитывать его на внутренние вершины

$$p(q|s) = \frac{\sum_{c \in C(s)} p(q|c)}{\sum_q \sum_{c \in C(s)} p(q|c)},$$

где q - состояние, $C(s)$ - все листья (контексты), являющиеся потомками s

Изначально берется полное дерево некоторой фиксированной глубины m .

Обучение модели по цепочки наблюдений происходит с помощью чередования ЕМ-алгоритма и обрезания дерева.

ЕМ-часть максимизирует правдоподобие модели алгоритмом Баума-Велха

Обрезание дерева производится нахождением родителей листьев, все потомки которых имеют близкие распределения переходов. Все дети такой вершины ликвидируются, а сама вершина становится новым контекстом. Критерием сравнения распределений переходов служит расстояние Кульбака-Лейблера.

Инициализация каждого следующего ЕМ ведется с помощью пересчета распределения переходов на новые контексты.

Изначальная инициализация производится путем построения цепочки состояний алгоритмом k -means по наблюдениям (где $k = n$) и частотной оценкой вероятности переходов на ней.

3 Оценка модели

4 Заключение

Список литературы

- [1] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.
- [2] Lawrence Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [3] Y Wang, Lizhu Zhou, and Jianhua Feng. Mining complex time-series data by learning Markovian models. In *International Conference on Data Mining*, 2006.