Санкт-Петербургский государственный университет
Прикладная математика и информатика

Отчет по Научно-исследовательской работе

Задачи оценивания значимости выравнивания при помощи скрытых марковских моделей

Выполнил:
Власенко Даниил Владимирович
группа 19.Б04-мм

Научный руководитель:
д. ф.-м. н., профессор
Коробейников Антон Иванович
Кафедра Статистического Моделирования

Санкт-Петербург
2022

1. Введение

Последовательность длины L — строка D состоящая из L символов алфавита Σ. Выравнивание последовательностей — размещение двух или более последовательностей друг под другом таким образом, чтобы было легче увидеть их схожие участки. Например, даны последовательности ACEAAFAE и CEAFDCE, если расположить их друг под другом, то не будет ни одного совпадения соответствующих символов, но если вставить пропуск восьмого символа в первой последовательности и пропуски первого и пятого символов во второй последовательности, то мы получим 5 совпадений:

A C E A A F A E
C E A F D C E

A C E A A F A — E
— C E A — F D C E

Рис. 1. Последовательности до и после выравнивания.

Определение 1. *Значимость выравнивания — действительное число s, отражающее сходство последовательностей.*

Способом вычисления значимости выравнивания s может быть, например, увеличение значимости на 1 при совпадении символов, стоящих друг под другом, и уменьшение на $\frac{1}{2}$ при несовпадении. Тогда значимость s приведенного выше выравнивания будет равна 3. Способ вычисления значимости выравнивания выбирается исходя из целей и вида выравнивания.

Сходство последовательностей может отражать функциональные, структурные или эволюционные взаимосвязи объектов, которые описывают эти последовательности. Таким образом вычисление значимости выравнивания последовательностей может быть полезно в задаче определения степени родства биологических организмов путем сравнения их ДНК или РНК, нуклеотидных последовательностей, задаче анализа свойств белков, аминокислотных последовательностей, задаче распознавания речи человека или письменного языка и многих других приложениях.

Выше был приведен пример попарного выравнивания двух строк, но если сходство последовательностей слабое, то через такое выравнивание может не выйти идентифицировать взаимосвязь описываемых последовательностями объектов. Однако сравнение сразу трех

и более последовательностей может позволить выявить эту взаимосвязь, такое выравнивание называется множественным. Проводить множественное выравнивание стандартными методами динамического программирования для попарного выравнивания [1] вычислительно неэффективно, но оказывается, что аппарат скрытых марковских моделей (СММ) позволяет эффективно решать эту задачу [2, 3].

СММ будут описаны далее, пока что зададимся следующим вопросом. Если есть множество последовательностей, описывающих взаимосвязанные объекты, имеется еще одна последовательность и была посчитана значимость выравнивания этой последовательности ко всему множеству каким-либо способом, то

- достаточно ли высокая эта значимость, чтобы считать объект, описываемый последовательностью, родственным к объектам, описываемым множеством, или шум, т.е. случайная последовательность, мог добиться такой значимости.
- достаточно ли низкая эта значимость, чтобы считать объект описываемый последовательностью, не родственным к объектам, описываемым множеством, или сигнал, т.е. последовательность, описывающая взаимосвязанный с множеством объект, мог получить такую значимость.

Определение 2. *Ложноположительная вероятность значимости s — это вероятность того, что шум получит значимость равную или выше s.*

Далее будет описаны метод, который позволяет эффективно вычислять введенный выше термин [4]. Исходный код можно изучить на GitHub [5].

2. Метод

Сначала опишем модели, затем алгоритмы, которые используются для манипуляции ими.

2.1. Модели

Метод предполагает, что даны профильная СММ [3], с помощью которой будут оцениваться последовательности, и фоновая модель B, которая будет описывать шум.

Скрытые марковские модели

Определение 3. *Пусть X_n и Y_n дискретные стохастические процессы, $n \geq 1$. Пара (X_n, Y_n) называется скрытой марковской моделью, если*

- X_n — *марковский процесс, поведение которого напрямую не наблюдается ("скрытый");*
- $\mathsf{P}(Y_n = y_n | X_1 = x_1, \ldots, X_n = x_n) = \mathsf{P}(Y_n | X_n = x_n)$ *для любого* $n \geq 1$, *где* $x_1, \ldots, x_n$ — *значения, принимаемые процессом* X_n (***состояния модели***), y_n — *значение, принимаемое процессом* Y_n (***наблюдаемый символ модели***).

Примером простой СММ может быть модель, изображенная на рисунке 1 и описывающая подбрасывание двух монет. Пусть между наблюдателем и человеком с монетами стоит ширма, которая позволяет наблюдателю видеть только пол, куда падают монеты. Пусть есть две монеты: одна — честная монета, вторая — нечестная монета с перевесом в одну из сторон. Пусть человек с монетами с некоторой вероятностью либо подбрасывает монету, которую он бросил в прошлый раз, либо меняет монеты и бросает новую. При этом наблюдатель не знает, какая монета используется в конкретный момент времени, так как он не видит рук бросающего монеты и не может отличить одну монету от другой по их внешнему виду, он видит только последовательность результатов бросков.

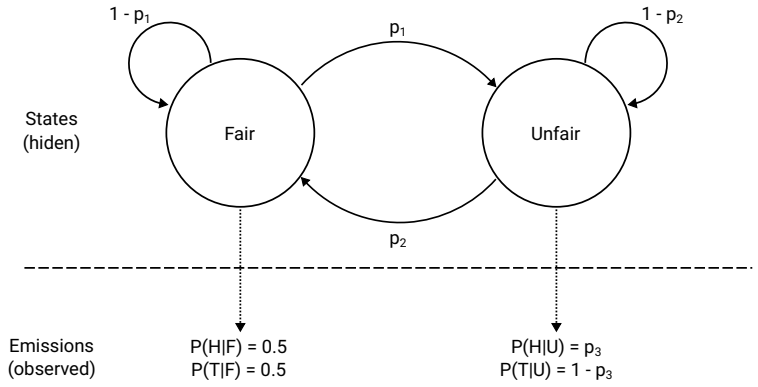

Рис. 2. Простая скрытая марковская модель.

Профильная скрытая марковская модель

Профильная СММ — это СММ со специальной линейной архитектурой состояний, которая позволяет выравнивать последовательность к множеству последовательностей.

Если для удобства реализации алгоритмов добавить специальное *начальное* и специальное *конечное* состояния, в которых профильная СММ начинает и заканчивает работу и не испускает наблюдаемых символов, как показано на рисунке 2, тогда *путь* π в профильной СММ начинается в начальном состоянии, заканчивается в конечном состоянии и проходит от состояния к состоянию, испуская в каждом состоянии наблюдаемый символ, то есть мы считаем, что путь π включает в себя и состояния, и наблюдаемые символы. *Последователь-*

ность D, связанная с путем π — последовательность наблюдаемых символов, которая была получена в результате прохода профильной СММ пути π.

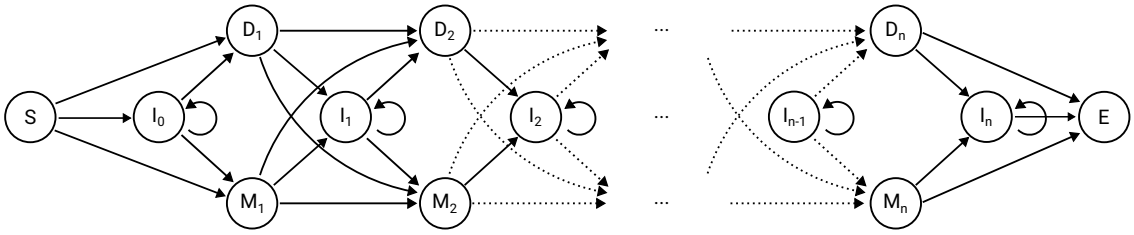

Рис. 3. Профильная скрытая марковская модель.

Вероятность пути $s(\pi)$ — произведение всех переходных вероятностей от состояний к состоянию и вероятностей наблюдаемых символов, которые излучаются в каждом состоянии, кроме начального и конечного, на протяжении всего пути π.

Вероятность последовательности D может интерпретироваться и считаться по-разному — алгоритмом *Витерби* или *Форвард* алгоритмом [2, 3].

Вероятность Витерби $s_{max}(D)$ последовательности D — это максимальная вероятность последовательности D среди всех путей π, которые могли бы ее испустить:

$$s_{max}(D) = \max_{\pi \in \pi_D}(s(\pi)).$$

Несмотря на большое количество возможных путей, которые могли бы испустить последовательность D, алгоритм Витерби позволяет эффективно решать эту задачу.

Форвард вероятность $s_{fw}(D)$ последовательности D — это общая вероятность того, что в результате работы СММ будет получена последовательность D:

$$s_{fw}(D) = \sum_{\pi \in \pi_D} s(\pi).$$

Форвард алгоритм работает за то же время, что и алгоритм Витерби.

Третий способ оценивать последовательности, позволяющий уменьшить дисперсию дальнейших вычислений оценки ложноположительной вероятности значимости, заключается в том, что каждая вероятность перехода из одного состояния в другое и вероятность излучения символа состоянием будут возводится в степень $\frac{1}{T}$, где $T \in (0; +\infty)$. При этом логика вычислений остается та же, то есть $s(\pi)^{\frac{1}{T}}$ и $s(D)^{\frac{1}{T}}$ будут вычисляться как вероятность произведения независимых событий и как сумма непересекающихся событий соответственно, хотя они уже могут не являться вероятностями (Например, сумма всех $s(\pi)^{\frac{1}{T}}$ не обязательно равна единице):

$$Z(D, T) = \sum_{\pi \in \pi_D} s(\pi)^{\frac{1}{T}}.$$

Функция $Z(D,T)$ называется статистической суммой и вычисляется через модификацию Форвард алгоритма. Метод подбора параметра T будет описан далее.

Фоновая модель

Мы предполагаем наличие простой фоновой модели B для последовательностей длины L такой, что все L символьных позиций независимы и одинаково распределены в соответствии с некоторым распределением $\mathsf{P}(d|B)$, где d отражает возможный наблюдаемый символ:

$$\mathsf{P}(D|B) = \prod_{i=1}^{L} \mathsf{P}(d_i|B), \tag{1}$$

где d_i — это i-ый наблюдаемый символ последовательности D.

2.2. Постановка математической проблемы

Вероятность последовательности D длины L сравнивается с остальными последовательностями той же длины. Определим ложноположительную вероятность значимости более строго.

Определение 4. *Ложноположительная вероятность значимости s_0 выравнивания последовательностей длины D определяется следующим уравнением:*

$$fpr(s_0) = \sum_{D \in D_L} \mathsf{P}(D|B)\Theta(s(D) \geq s_0), \tag{2}$$

где $\mathsf{P}(D|B)$ — условная вероятность последовательности D, описываемая фоновой моделью, $s(D)$ — вероятность последовательности D, считаемая профильной CMM, и

$$\Theta(s(D) \geq s_0) = \begin{cases} 1, & s(D) \geq s_0 \\ 0, & s(D) < s_0 \end{cases}.$$

То есть $fpr(s_0)$ — это вероятность того, что шум достигнет или превзойдет значимость s_0. В определении $fpr(s_0)$ вероятность D отмечена как $s(D)$, потому что способ оценки последовательности может выбираться относительно интересующего приложения, подходит $s(D) = s_{max}(D)$ и $s(D) = s_{fw}(D)$.

2.3. Алгоритм

Выборка по значимости

Так как вычисление $fpr(s_0)$ через формулу (2) обычно неосуществимо, значение $fpr(s_0)$ может быть оценено через выборку по значимости, то есть через моделирование строк в

соответствии с фоновой моделью B и оценивание значения $fpr(s_0)$ долей тех из них, что достигают значимости s_0.

Построим распределение, относительно которого будем моделировать строки. Пусть $\mathsf{P}(D|T)$ — это условная вероятность строки D относительно некоторой модели строк длины L параметризованной значением T. Тогда можно переписать $fpr(s_0)$:

$$fpr(s_0) = \sum_{D \in D_L} \mathsf{P}(D|T) f(D, s_0),$$

где

$$f(D, s_0) = \frac{\mathsf{P}(D|B)\Theta(s(D) \geq s_0)}{\mathsf{P}(D|T)}. \tag{3}$$

Мы можем оценить значение $fpr(s_0)$ через моделирование последовательностей в соответствии с этой альтернативной моделью и подсчет среднего значения $f(D, s_0)$. Этот подход и называется *выборкой по значимости*, он полезен, потому что если правильно подобрать альтернативную модель, то удастся уменьшить дисперсию оценки $fpr(s_0)$:

$$\sum_{D \in D_L} \mathsf{P}(D|T)(f(D, s_0) - fpr(s_0))^2 \ll \sum_{D \in D_L} \mathsf{P}(D|B)(\Theta(s(D) \geq s_0) - fpr(s_0))^2.$$

Определим модель, используемую для выборки по важности параметризованную T следующим образом:

$$\mathsf{P}(D|T) = \frac{\mathsf{P}(D|B)Z(D, T)}{Z(T)},$$

где

$$Z(T) = \sum_{D \in D_L} \mathsf{P}(D|B) Z(D, T). \tag{4}$$

Подставив определение $\mathsf{P}(D, T)$ в уравнение (3) получим

$$f(D, s_0) = \frac{Z(T)\Theta(s(D) \geq s_0)}{Z(D, T)}.$$

Моделирование выборки

В итоге мы хотим смоделировать последовательности в соответствии с распределением $\mathsf{P}(D|T)$, вычислить $f(D, s_0)$ для каждой последовательности и использовать среднее этих значений как оценку $fpr(s_0)$. Здесь будет описан метод моделирования последовательностей.

Сначала, используя фоновую модель определенную уравнением (1), вычисляется значение $Z(D)$ через модификацию Форвард алгоритма, вычисляющего $Z(D, T)$. В алгоритме, вычисляющем $Z(D, T)$, излучение символа d некоторым состоянием E связывалось с вероятностью излучения этого символа этим состоянием, возведенной в степень $\frac{1}{T}$ — $s_E(d)^{\frac{1}{T}}$. В

алгоритме вычисляющем $Z(T)$ вместо такого множителя используется среднее значение излучений для состояния E:

$$\langle s_E^T \rangle_B = \sum_{d'} \mathsf{P}(d'|B) s_E(d)^{\frac{1}{T}}.$$

Потому что намного эффективнее заранее вычислить эти значения и хранить их, чем вычислять значение $Z(T)$ напрямую через формулу (4).

Мы моделируем строку длины L обратным ходом по форвард таблице, полученной в результате вычисления $Z(T)$. А точнее мы моделируем путь π, при этом вероятность излучения состоянием E символа d' следующая:

$$\mathsf{P}_E(d') = \frac{\mathsf{P}(d'|B) s_E(d')^{\frac{1}{T}}}{\langle s_E^T \rangle_B}.$$

Таким образом мы моделируем путь π из распределения

$$\mathsf{P}(\pi|T) = \frac{\mathsf{P}(D|B) s(\pi)^{\frac{1}{T}}}{Z(T)}.$$

Дальше мы оставляем только наблюдаемые символы, забывая состояния, и получаем строку D. Так как строка D могла быть излучена разными путями, получаем следующую вероятность моделирования строки D этим методом:

$$\mathsf{P}(D|T) = \sum_{\pi \in \pi_D} \frac{\mathsf{P}(D|B) s(\pi)^{\frac{1}{T}}}{Z(T)}.$$

Оценивание ложноположительной вероятности значимости

Мы хотим оценить ложноположительную вероятность выравнивания для значения s_0. Для каждой последовательности из N смоделированных последовательностей $\{D_i : i = 1, \ldots, N\}$ мы вычисляем $s(D_i)$ и $Z(D_i, T)$. Тогда оценка $fpr(s_0)$ следующая:

$$\widehat{fpr}(s_0) = \frac{Z(T)}{N} \sum_1^N \frac{\Theta(s(D_i) \geq s_0)}{Z(D_i, T)} = 1 - \widehat{tpr}(s_0),$$

где $\widehat{tpr}(s_0)$ оценка истиннонегативной вероятности выравнивания s_0.

Выбор T

Так как связь между параметром T и значимостью выравнивания s_0 неявная, то перед осуществлением алгоритма нужно сначала проверить, при каком параметре T оценка $fpr(s_0)$ имеет меньшую дисперсию. Для этого придется смоделировать несколько строк для разных T, и только потом моделировать выборку из N строк с подходящим параметром T. На практике было замечена меньшая дисперсия у таких T, при которых 20–60% смоделированных строк удовлетворяют неравенству $s(D) \geq s_0$.

3. Результаты

Для построения профильной СММ необходимо иметь выравнивание последовательностей, которые считаются взаимосвязанными. При этом если в какой-либо колонке встречается большое количество пропусков, то эта колонка с большей вероятностью может не отражать качества, свойственные всему множеству объектов, описываемых последовательностями. Такие колонки отмечаются и обрабатываются особым образом при построении профильной СММ. Долю же пропусков, которая необходима для того, чтобы считать колонку неважной, определяют в зависимости от решаемой задачи. Пусть в нашем случае эта доля будет равна $\frac{2}{5}$.

Будем использовать выравнивание пяти последовательностей, изображенное на рисунке 4. Шестая и седьмая колонки прозрачнее остальных тем самым отмечены как неважные.

A C D E F A C A D F
A F D A — — — C C F
A — — E F D — F D C
A C A E F — — A — C
A D D E F A A A D F

Рис. 4. Пример множественного выравнивания.

Нас интересует значение ложноположительной вероятности значимостей $s_0 = 10^{-3}$, $s_0 = 10^{-6}$ и $s_0 = 10^{-9}$ для строк длины $L = 8$.

Для каждого значения было смоделировано 50 последовательностей для различных параметров T, чтобы отобрать подходящие значения параметра. Далее для подобранных T для каждого из четырех значений s_0 была смоделирована выборка из 1000 последовательностей и выполнена оценка $\widehat{fpr}(s_0)$ через описанный выше алгоритм, в котором за вероятность последовательности $s(D)$ была взята Форвард вероятность $s_{fw}(D)$. Результаты можно наблюдать в таблице 1, в которой так же указано настоящее значение $fpr(s_0)$, полученное по формуле (2), и доверительный интервал уровня $\gamma = 0.99$. Как можно наблюдать чем выше значимость s_0 тем меньше вероятность того, что шум может достичь такого значения.

Результаты вычислений оценки $\widehat{fpr}(s_0)$ для строк длины $L = 100$ и доверительные интервалы уровня $\gamma = 0.99$ изображены в таблице 2.

Таблица 1. Результаты для коротких строк

s_0	T	$\widehat{fpr}(s_0)$	$fpr(s_0)$	$[c_1(\gamma); c_2(\gamma)]$
10^{-3}	1	0.000130127	0.00013312	[0.0000117484; 0.0008517416]
10^{-6}	2	0.0105416	0.0102349	[0.0081853; 0.0134781]
10^{-9}	3	0.214698	0.21278	[0.20366; 0.22479]

Таблица 2. Результаты для длинных строк

s_0	T	$\widehat{fpr}(s_0)$	$[c_1(\gamma); c_2(\gamma)]$
10^{-85}	7	0.0000000183	[0.0; 0.00066349]
10^{-90}	7	0.003175	[0.001884; 0.004779]
10^{-100}	7	0.615709	[0.597540 0.622677]

Список литературы

1. Compeau Phillip, Pevzner Pavel. How do we compare DNA sequences // Bioinformatics Algorithms: An Active Learning Approach, 2nd Ed. Vol. 1. — Active Learning Publishers, 2015.
2. A tutorial on Hidden Markov Models : Rep. / Signal Processing and Artifcial Neural Networks Laboratory Department of Electrical Engineering Indian Institute of Technology ; Executor: Rakesh Dugad, U. B. Desai : 1996.
3. Compeau Phillip, Pevzner Pavel. Why have biologists still not developed an HIV vaccine // Bioinformatics Algorithms: An Active Learning Approach, 2nd Ed. Vol. 2. — Active Learning Publishers, 2015.
4. Newberg Lee A. Error statistics of hidden Markov model and hidden Boltzmann model results // MC Bioinformatics. — 2009.
5. Vlasenko Daniil. Significance estimation. — https://github.com/Daniil-Vlasenko/SPbUSignificanceEstimation. — 2022.
6. Stamp Mark. Introduction to Machine Learning with Applications in Information Security // A revealing introduction to Hidden Markov Models. — Chapman and Hall, 2021.
7. Jurafsky Daniel, Martin James H. Speech and Language Processing // Hidden Markov Models. — Prentice Hall, 2021.
8. Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition // Proceedings of the IEEE. — 1989. — P. 257 – 286.
9. Newberg Lee A. Significance of Gapped Sequence Alignments // Journal of Computational Biology. — 2008.