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

Оглавление

1. Введение

Пусть дан алфавит символов Σ.

Определение 1. *Последовательностью длины L над алфавитом Σ будем называть такой X, что $X \in \Sigma^L$. Последовательностью X над алфавитом Σ будем называть такой X, что $X \in \bigcup_{L=0}^{L=\infty} \Sigma^L$.*

Сходство последовательностей может отражать функциональные, структурные или эволюционные взаимосвязи объектов, которые описывают эти последовательности. Таким образом умение находить взаимосвязи в строках может быть приложимо 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.

Определение 2. *Выравниванием N последовательностей называется отображение $Q : \times_{i=1}^{N}(\bigcup_{L_i=0}^{\infty} \Sigma^{L_i}) \to \times_{i=1}^{N}(\Sigma^L)$, где $L = \max(\{L_i\}_{i=1}^{N})$ такое что:*

1. *Возможны вставки символа — в последовательностях.*
2. *Вставка — на одинаковых позициях во всех последовательностях запрещена.*
3. *Порядок изначальных символов внутри последовательностей сохраняется.*

Элементы из области значения Q также называются выравниваниями. Примем множество $\times_{i=1}^{N}(\bigcup_{L_i=0}^{L_i=\infty} \Sigma^{L_i})$ за пространство элементарных исходов Ω. Область значений выравнивания Q обозначим как $\overline{\Omega}$.

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 3. *Оценкой выравнивания называется случайная величина $s : \overline{\Omega} \to \mathbb{R}$.*

Способом вычисления оценки выравнивания s может быть, например, увеличение оценки на 1 при совпадении символов, стоящих на одинаковых позициях в последовательностях, и уменьшение на $\frac{1}{2}$ при несовпадении. Тогда оценка s приведенного на рисунке 1 выравнивания будет равна 3.

Определить оценку выравнивания можно разными способами, но смысл будет иметь такое определение, чтобы оценка была мерой того, насколько сильно строки выравнивания похожи друга на друга.

Пусть даны последовательности $\{X_i\}_{i=1}^N$ и задана оценка выравнивания s. Тогда задача оценки сходства последовательностей $\{X_i\}_{i=1}^N$ сводится к решению оптимизационной задачи:

$$\max_{\overline{\omega}\in\overline{\Omega}:Q(\{X_i\}_{i=1}^N)=\overline{\omega}} s(\overline{\omega}).$$

Предположим, что даны последовательность X и выравнивание $\omega \in \overline{\Omega}$ из N строк.

Определение 4. *Выравниванием последовательности X к выравниванию w называется отображение $\boldsymbol{Q} : (X, \overline{\Omega}) \to \times_{i=1}^{N+1}(\Sigma^L)$, где $L = \max(\{L_i\}_{i=1}^{N+1})$ такое что:*

1. *Возможны вставки символа — в последовательностях.*
2. *Вставка — на одинаковых позициях во всех последовательностях запрещена.*
3. *Порядок изначальных символов внутри последовательностей сохраняется.*

Примем множество $(\Omega, \overline{\Omega})$ за пространство элементарных исходов $\boldsymbol{\Omega}$. Область значений выравнивания $\boldsymbol{Q}$ обозначим как $\overline{\boldsymbol{\Omega}}$.

Определение 5. *Оценкой выравнивания последовательности X к выравниванию w называется случайная величина $\boldsymbol{s} : \overline{\boldsymbol{\Omega}} \to \mathbb{R}$.*

Пусть даны последовательность X, выравнивание $\overline{\omega} \in \overline{\Omega}$ и задана оценка выравнивания $\boldsymbol{s}$. Тогда задача оценки сходства последовательности X и последовательностей, описываемых выравниванием $\overline{\omega}$, сводится к решению оптимизационной задачи:

$$\max_{\overline{\boldsymbol{\omega}}\in\overline{\boldsymbol{\Omega}}:\boldsymbol{Q}(X,\overline{\omega})=\overline{\boldsymbol{\omega}}} s(\overline{\boldsymbol{\omega}}).$$

Встает вопрос того, как интерпретировать решение этой задачи. Пусть дана последовательность X, выравнивание $\overline{\omega} \in \overline{\Omega}$, задана оценка оценка выравнивания $\boldsymbol{s}$ и известно, что $\overline{\omega}$ построено на последовательностях, описывающих взаимосвязанные объекты.

Определение 6. *Шумом будем называть случайную последовательность над алфавитом Σ. Сигналом будем называть последовательность над алфавитом Σ, которая описывает объект, взаимосвязанный с объектами последовательностей, описываемых $\overline{\omega}$.*

Достаточно ли низкая оценка $\boldsymbol{s}(X, \overline{\omega})$, чтобы считать последовательность X шумом, или сигнал мог получить такую оценку?

2. Метод

Для ответа на поставленный вопрос сначала опишем нужные нам модели, затем алгоритмы, которые используются для манипуляции ими.

2.1. Модели

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

2.1.1. Скрытая марковская модель

Определение 7. *Пусть Z_n и Y_n дискретные стохастические процессы, $n \geqslant 1$. Пара (Z_n, Y_n) называется скрытой марковской моделью, если*

- *Z_n — марковский процесс, поведение которого напрямую не наблюдается ("скрытый");*
- *$P(Y_n = y_n | Z_1 = z_1, \ldots, Z_n = z_n) = P(Y_n | Z_n = z_n)$ для любого $n \geqslant 1$, где $z_1, \ldots, z_n$ — значения, принимаемые процессом Z_n (**состояния модели**), y_n — значение, принимаемое процессом Y_n (**наблюдаемый символ модели**).*

Определение 8. *Путем π называется последовательность состояний $\{z_i\}_{i=1}^n$ и наблюдаемых символов $\{y_i\}_{i=1}^n$ СММ. Последовательность X, которая была получена в результате прохода профильной СММ пути π, называется последовательностью наблюдаемых символов.*

Примером простой СММ может быть модель, изображенная на рисунке 2 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.

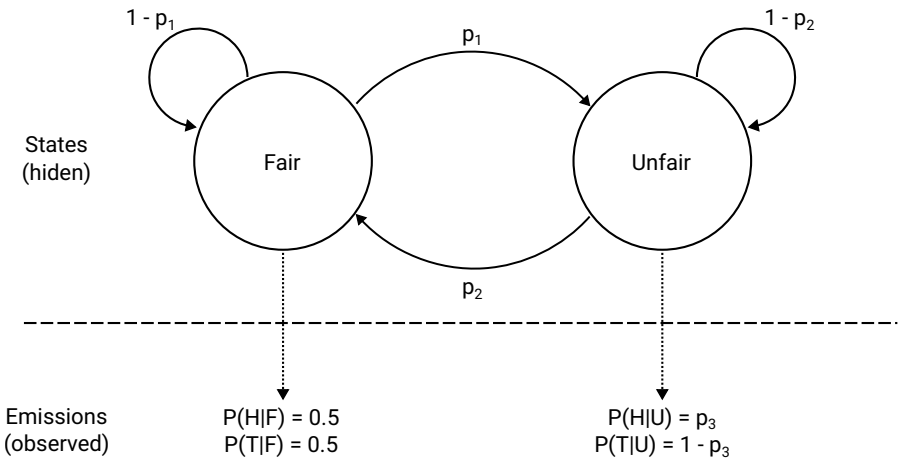

Рис. 2. Простая скрытая марковская модель.

2.1.2. Профильная скрытая марковская модель

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

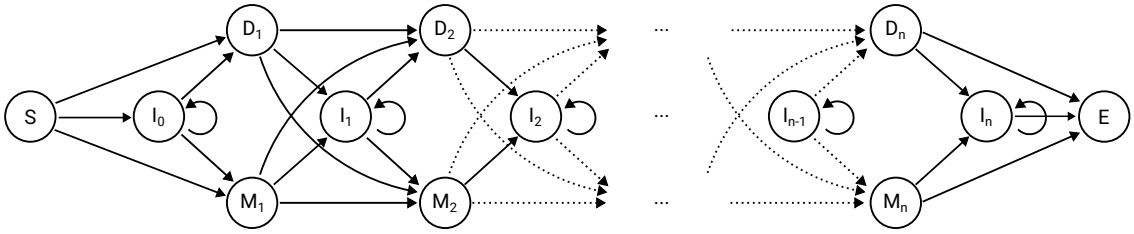

Рис. 3. Профильная скрытая марковская модель.

Пусть даны последовательность X и выравнивание $\overline{\omega} \in \overline{\Omega}$. Профильная CMM состоит из трех типов состояний, распределения которых строятся на основе $\overline{\omega}$:

- S-состояние — начальное состояние,
- M-состояния — устанавливают соответствие символов в X и $\overline{\omega}$,
- I-состояния и D-состояния — устанавливает соответствие пропуска и символа в X и $\overline{\omega}$,
- E-состояние — конечное состояние.

Алгоритмы профильных CMM позволяют по-разному оценивать выравнивание X к $\overline{\omega}$.

Определение 9. *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 Витерби $s_{max}(X)$ последовательности X — это максимальная вероятность последовательности X среди всех путей π, которые могли бы ее испустить:

$$s_{max}(X) = \underset{\pi \in \pi_X}{max}(s(\pi)),$$

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

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

$$s_{fw}(X) = \sum_{\pi \in \pi_X} s(\pi).$$

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

Третий способ оценивать последовательности, позволяющий уменьшить дисперсию дальнейших вычислений оценки ложноположительной вероятности оценки, 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 $s(X)^{\frac{1}{T}}$ 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):

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

Функция $Z(X,T)$ называется статистической суммой и вычисляется через модификацию Форвард алгоритма. Параметра T подбирается экспериментально под конкретную интересующую оценку выравнивания.

2.1.3. Фоновая модель

Определение 10. *Моделью последовательностей называется генератор, моделирующий последовательности в соответствии с некоторым распределением.*

$P(X|M)$, где M — некоторая модель, означает условную вероятность X, при условии ее моделирования моделью M.

Определение 11. *Фоновой моделью B для последовательностей длины L называется генератор последовательностей длины L такой, что все L символьных позиций независимы и одинаково распределены:*

$$\mathsf{P}(X|B) = \prod_{i=1}^{L} \mathsf{P}(x_i|B),$$

где x_i отражает возможный наблюдаемый символ.

Фоновая модель описывает шум.

2.1.4. Постановка математической задачи.

Последовательность X длины L сравнивается с остальными последовательностями той же длины.

Определение 12. *Ложноположительная вероятность оценки s_0 для строк длины L:*

$$fpr(s_0) = \sum_{X \in X_L} \mathsf{P}(X|B)\Theta(s(X) \geqslant s_0),$$

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

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

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

2.2. Алгоритм

2.2.1. Выборка по значимости

Так как вычисление $fpr(s_0)$ по определению 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 оценки s_0.

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

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

где

$$f(X, s_0) = \frac{\mathsf{P}(X|B)\Theta(s(X) \geqslant s_0)}{\mathsf{P}(X|T)}.$$

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

Определим распределение модели, используемой для выборки по важности параметризованную T следующим образом:

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

где

$$Z(T) = \sum_{X \in X_L} \mathsf{P}(X|B)Z(X,T).$$

Подставив определение $\mathsf{P}(X|T)$ в определение $f(X, s_0)$, получим

$$f(X, s_0) = \frac{Z(T)\Theta(s(X) \geqslant s_0)}{Z(X|T)}.$$

2.2.2. Моделирование выборки

В итоге мы хотим смоделировать последовательности в соответствии с распределением $\mathsf{P}(X|T)$, вычислить $f(X, s_0)$ 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 (11), вычисляется значение $Z(X)$ через модификацию Форвард алгоритма, вычисляющего $Z(X, T)$. В алгоритме, вычисляющем $Z(X, T)$, излучение символа x некоторым состоянием z связывалось с вероятностью излучения этого символа этим состоянием, возведенной в степень $\frac{1}{T} - s_z(x)^{\frac{1}{T}}$. В алгоритме вычисляющем $Z(T)$ вместо такого множителя используется среднее значение излучений для состояния z:

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

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

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 z символа x' следующая:

$$\mathsf{P}_z(x') = \frac{\mathsf{P}(x'|B) s_z(x')^{\frac{1}{T}}}{\langle s_z^T \rangle_B}.$$

Таким образом мы моделируем путь π из распределения

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

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

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

2.2.3. Оценивание ложноположительной вероятности значимости

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

$$\widehat{fpr}(s_0) = \frac{Z(T)}{N} \sum_{1}^{N} \frac{\Theta(s(X_i) \geqslant s_0)}{Z(X_i, T)}.$$

2.2.4. Выбор T

Так как связь между параметром T и оценкой 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 $s(X) \geqslant s_0$.

3. Результаты

Для построения профильной СММ необходимо иметь выравнивание последовательностей таких, что априорно известно, что они описывают взаимосвязанные объекты. 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.

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]

Нас интересует значение ложноположительной вероятности оценок $s_0 = 10^{-3}$, $s_0 = 10^{-6}$ и $s_0 = 10^{-9}$ в смысле Форвад вероятностей 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 (12), 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 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.