Задачи оценивания значимости выравнивания при помощи скрытых марковских моделей

Власенко Даниил Владимирович, гр.19.Б04-мм

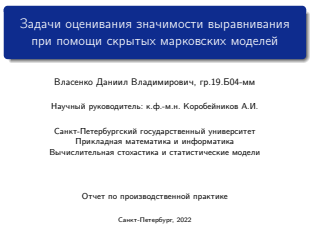
Научный руководитель: к.ф.-м.н. Коробейников А.И.

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Отчет по производственной практике

Санкт-Петербург, 2022

Оценивание значимости выравнивания



Научный руководитель к.ф.-м.н., Коробейников А.И.,
кафедра статистического моделирования

Введение

Пусть дан алфавит символов Σ.

Определение

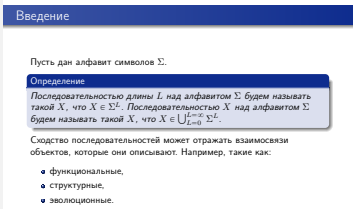
Последовательностью длины L над алфавитом Σ будем называть такой X, что $X \in \Sigma^L$. Последовательностью X над алфавитом Σ будем называть такой X, что $X \in \bigcup_{L=0}^{L=\infty} \Sigma^L$.

Сходство последовательностей может отражать взаимосвязи объектов, которые они описывают. Например, такие как:

- функциональные,
- структурные,
- эволюционные.

Оценивание значимости выравнивания

└Введение



Пусть дан алфавит символов Σ.

Определение

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.

Введение

Определение

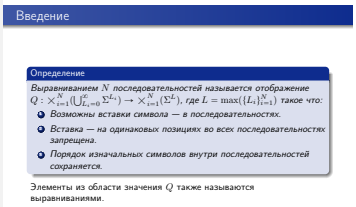
Выравниванием N последовательностей называется отображение $Q : \times_{i=1}^{N}(\bigcup_{L_i=0}^{\infty} \Sigma^{L_i}) \to \times_{i=1}^{N}(\Sigma^{L})$, где $L = \max(\{L_i\}_{i=1}^{N})$ такое что:

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.

Оценивание значимости выравнивания

└Введение



Определение

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.

Введение

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.

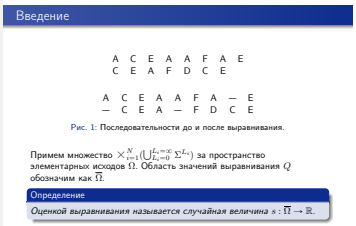
Примем множество $\times_{i=1}^{N}(\bigcup_{L_i=0}^{L_i=\infty}\Sigma^{L_i})$ за пространство элементарных исходов Ω. Область значений выравнивания Q обозначим как $\overline{\Omega}$.

Определение

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

Оценивание значимости выравнивания

└Введение



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 $\overline{\Omega}$.

Определение

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 слайде 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.

Введение

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

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

Оценивание значимости выравнивания

└Введение

Введение

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

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

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:

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

Введение

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

Определение

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.*

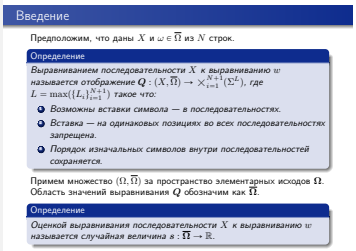
Примем множество $(\Omega, \overline{\Omega})$ за пространство элементарных исходов $\boldsymbol{\Omega}$. Область значений выравнивания $\boldsymbol{Q}$ обозначим как $\overline{\boldsymbol{\Omega}}$.

Определение

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

Оценивание значимости выравнивания

└Введение



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

Определение

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

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

Введение

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

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

Оценивание значимости выравнивания

└Введение

Введение

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

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

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 $\overline{\omega}$, сводится к решению оптимизационной задачи:

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

Введение

Пусть дана $X \in \Sigma$, $\overline{\omega} \in \overline{\Omega}$, $\boldsymbol{s}$ и известно, что $\overline{\omega}$ построено на последовательностях, описывающих взаимосвязанные объекты.

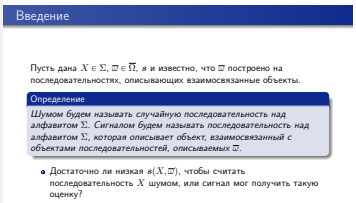
Определение

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

Оценивание значимости выравнивания

└Введение



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

Определение

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?

Обозначения и известные результаты

Определение

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**).*

Оценивание значимости выравнивания

└Обозначения и известные результаты

Обозначения и известные результаты

Определение

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**).*

Для ответа на этот вопрос сначала опишем нужные нам модели, затем алгоритмы, которые используются для манипуляции ими.
Метод предполагает, что даны профильная СММ, с помощью которой будут оцениваться последовательности, и фоновая модель B, которая будет описывать шум.

Определение

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

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.

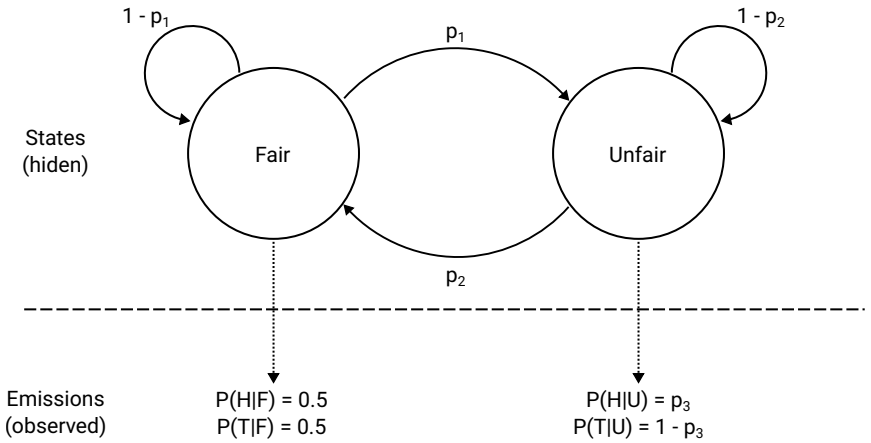
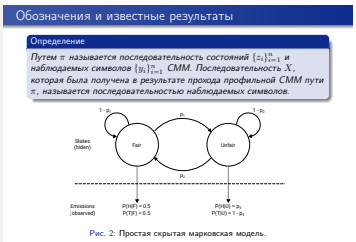


Рис. 2: Простая скрытая марковская модель.

Оценивание значимости выравнивания

└Обозначения и известные результаты



Определение

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 слайде 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.

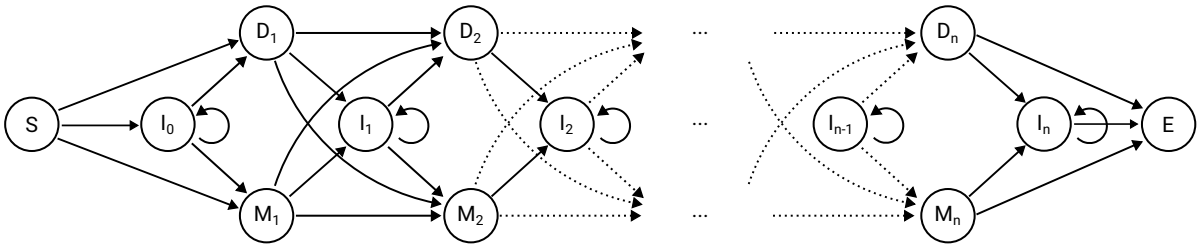


Рис. 3: Профильная скрытая марковская модель.

Пусть даны X, $\overline{\omega} \in \overline{\Omega}$. Профильные СММ состоят из трех типов состояний, распределения которых строятся на основе $\overline{\omega}$ (Compeau Phillip, Pevzner Pavel 2015):

- 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.

Оценивание значимости выравнивания

└Обозначения и известные результаты

Обозначения и известные результаты

Рис. 3: Профильная скрытая марковская модель.

Пусть даны X, $\overline{\omega} \in \overline{\Omega}$. Профильные СММ состоят из трех типов состояний, распределения которых строятся на основе $\overline{\omega}$ (Compeau Phillip, Pevzner Pavel 2015):

- 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.

Профильная СММ — это СММ со специальной линейной архитектурой состояний, которая позволяет выравнивать последовательность к множеству последовательностей.
Пусть даны X, $\overline{\omega} \in \overline{\Omega}$. Профильные СММ состоят из трех типов состояний, если не считать начальное и конечное, 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.

Обозначения и известные результаты

Алгоритмы профильных CMM позволяют по-разному оценивать выравнивание X к $\overline{\omega}$ (Newberg Lee A. 2009; Rakesh Dugad 1996).

Определение

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.

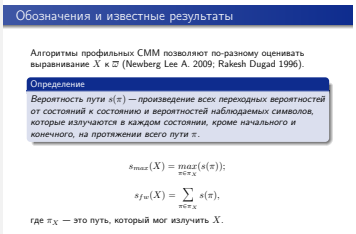
$$s_{max}(X) = \max_{\pi \in \pi_X}(s(\pi));$$

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

где π_X — это путь, который мог излучить X.

Оценивание значимости выравнивания

└Обозначения и известные результаты



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

Определение

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.

$$s_{max}(X) = \max_{\pi \in \pi_X}(s(\pi));$$

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

где π_X — это путь, который мог излучить X.

Обозначения и известные результаты

Алгоритмы профильных CMM позволяют по-разному оценивать выравнивание X к $\overline{\omega}$ (Newberg Lee A. 2009; Rakesh Dugad 1996).

Определение

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.

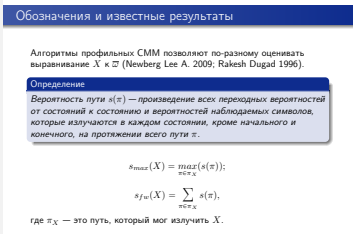
$$s_{max}(X) = \max_{\pi \in \pi_X}(s(\pi));$$

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

где π_X — это путь, который мог излучить X.

Оценивание значимости выравнивания

└Обозначения и известные результаты


Обозначения и известные результаты

Алгоритмы профильных CMM позволяют по-разному оценивать выравнивание X к $\overline{\omega}$ (Newberg Lee A. 2009; Rakesh Dugad 1996).

Определение

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.

$s_{max}(X) = \max_{\pi \in \pi_X}(s(\pi));$

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

где π_X — это путь, который мог излучить X.

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

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

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 CMM 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.

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

где $s(\pi)^{\frac{1}{T}}$ обозначает эквивалент вероятности π, который все еще вычисляется как произведение независимых событий, но каждый множитель возводиться в степень $\frac{1}{T}$.

Оценивание значимости выравнивания

└Обозначения и известные результаты

Обозначения и известные результаты

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

где $s(\pi)^{\frac{1}{T}}$ обозначает эквивалент вероятности π, который все еще вычисляется как произведение независимых событий, но каждый множитель возводиться в степень $\frac{1}{T}$.

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.

Обозначения и известные результаты

Определение

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

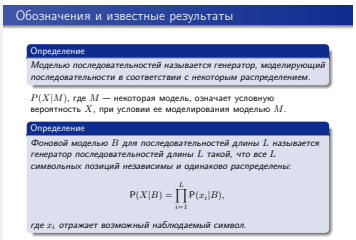
Фоновой моделью B для последовательностей длины L называется генератор последовательностей длины L такой, что все L символьных позиций независимы и одинаково распределены:

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

где x_i отражает возможный наблюдаемый символ.

Оценивание значимости выравнивания

└Обозначения и известные результаты



Определение

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

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

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

где x_i отражает возможный наблюдаемый символ.

Фоновая модель описывает шум.

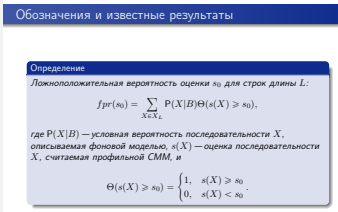
Определение

Ложноположительная вероятность оценки s_0 для строк длины L:

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

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

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



Последовательность X длины L сравнивается с остальными последовательностями той же длины. Определим ложноположительную вероятность оценки:

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

где $\mathrm{P}(X|B)$ — 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 $s(X) = s_{fw}(X)$.

Обозначения и известные результаты

Вычисление $fpr(s_0)$ по определению обычно неосуществимо, значение $fpr(s_0)$ может быть оценено через выборку по значимости.

Пусть $\mathrm{P}(X|T)$ — это условная вероятность строки X относительно некоторой модели строк длины L параметризованной значением T. Тогда можно переписать $fpr(s_0)$:

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

где

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

Оценивание значимости выравнивания

└Обозначения и известные результаты

Обозначения и известные результаты

Вычисление $fpr(s_0)$ по определению обычно неосуществимо, значение $fpr(s_0)$ может быть оценено через выборку по значимости.

Пусть $\mathrm{P}(X|T)$ — это условная вероятность строки X относительно некоторой модели строк длины L параметризованной значением T. Тогда можно переписать $fpr(s_0)$:

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

где

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

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 $\mathrm{P}(X|T)$ — это условная вероятность строки X относительно некоторой модели строк длины L параметризованной значением T. Тогда можно переписать $fpr(s_0)$:

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

где

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

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 $fpr(s_0)$.

Обозначения и известные результаты

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

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

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

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

Оценивание значимости выравнивания

└Обозначения и известные результаты

Обозначения и известные результаты

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

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

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

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

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

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

Обозначения и известные результаты

Моделируем $\{X_i\}_{i=1}^N$ в соответствии с распределением $\mathrm{P}(X|T)$ (Newberg Lee A. 2009), вычисляем $f(X, s_0)$ для каждой последовательности и использовать среднее этих значений как оценку $fpr(s_0)$:

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

Оценивание значимости выравнивания

└Обозначения и известные результаты

Обозначения и известные результаты

Моделируем $\{X_i\}_{i=1}^N$ в соответствии с распределением $\mathrm{P}(X|T)$ (Newberg Lee A. 2009), вычисляем $f(X, s_0)$ для каждой последовательности и использовать среднее этих значений как оценку $fpr(s_0)$:

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

Моделируем $\{X_i\}_{i=1}^N$ в соответствии с распределением $\mathrm{P}(X|T)$, вычисляем $f(X, s_0)$ для каждой последовательности и использовать среднее этих значений как оценку $fpr(s_0)$:

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

Опуская подробности того как устроены моделирование, построенное на модификациях классических алгоритмов, связанных с HMM, и метод подбора параметра T, перейдем к полученным результатам.

Полученные результаты

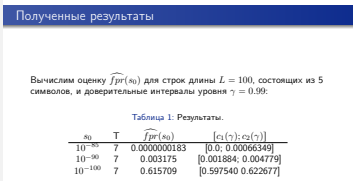
Вычислим оценку $\widehat{fpr}(s_0)$ для строк длины $L = 100$, состоящих из 5 символов, и доверительные интервалы уровня $\gamma = 0.99$:

Таблица 1: Результаты.

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]

Оценивание значимости выравнивания

└Полученные результаты



Вычислим оценку $\widehat{fpr}(s_0)$ для строк длины $L = 100$ и доверительные интервалы уровня $\gamma = 0.99$:

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]

Вычисление $fpr(s_0)$ перебором привело бы к перебору 5^{100} строк.

Заключение

- Была изучена тема алгоритмов парного и множественного выравнивания последовательностей и тема CMM и алгоритмов взаимодействия с ними.
- Был реализован алгоритм, позволяющий эффективно вычислять оценку $fpr(s_0)$.
- Предстоит подробно верифицировать реализованный алгоритм и сравнить его с имеющимися методами вычисления оценки $fpr(s_0)$.

Заключение

- Была изучена тема алгоритмов парного и множественного выравнивания последовательностей и тема CMM и алгоритмов взаимодействия с ними.
- Был реализован алгоритм, позволяющий эффективно вычислять оценку $fpr(s_0)$.
- Предстоит подробно верифицировать реализованный алгоритм и сравнить его с имеющимися методами вычисления оценки $fpr(s_0)$.

- Была изучена тема алгоритмов парного и множественного выравнивания последовательностей и тема CMM и алгоритмов взаимодействия с ними.
- Был реализован алгоритм, позволяющий эффективно вычислять оценку $fpr(s_0)$.
- Предстоит подробно верифицировать реализованный алгоритм и сравнить его с имеющимися методами вычисления оценки $fpr(s_0)$.