

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

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

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

Пусть дан алфавит символов Σ .

Определение

Последовательностью длины L над алфавитом Σ будем называть такой элемент 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

Введение

Пусть дан алфавит символов Σ .

Определение

Последовательностью длины L над алфавитом Σ будем называть такой элемент 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.

Пусть дан алфавит символов Σ .

Определение

Последовательностью длины L над алфавитом Σ будем называть такой элемент 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

Выравниванием N последовательностей называется отображение $Q : \times_{i=1}^N (\bigcup_{L_i=0}^{L_i=\infty} \Sigma^{L_i}) \rightarrow \times_{i=1}^N (\Sigma^{\max_{L \in L_i} (L)})$, 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

Введение

Определение

Выравниванием N последовательностей называется отображение $Q : \times_{i=1}^N (\bigcup_{L_i=0}^{L_i=\infty} \Sigma^{L_i}) \rightarrow \times_{i=1}^N (\Sigma^{\max_{L \in L_i} (L)})$, 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.

Определение

Выравниванием N последовательностей называется отображение $Q : \times_{i=1}^N (\bigcup_{L_i=0}^{L_i=\infty} \Sigma^{L_i}) \rightarrow \times_{i=1}^N (\Sigma^{\max_{L \in L_i} (L)})$, 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.

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 $\bar{\Omega}$.

Определение

Оценкой выравнивания называется случайная величина $s : \bar{\Omega} \rightarrow \mathbb{R}$.

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 $\bar{\Omega}$.

Определение

Оценкой выравнивания называется случайная величина $s : \bar{\Omega} \rightarrow \mathbb{R}$.

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 друг на друга.

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 $\bar{\Omega}$.

Определение

Оценкой выравнивания называется случайная величина $s : \bar{\Omega} \rightarrow \mathbb{R}$.

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:

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

Оценивание значимости выравнивания

└ Введение

Введение

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:

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

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:

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

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

Определение

Выравниванием последовательности X к выравниванию w называется отображение $Q : (X, \bar{\Omega}) \rightarrow \times_{i=1}^{N+1} (\Sigma^{\max_{L \in L_i} (L)})$, 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 $(\Omega, \bar{\Omega})$ за пространство элементарных исходов Ω . Область значений выравнивания Q обозначим как $\bar{\Omega}$.

Определение

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

Оценивание значимости выравнивания

— Введение

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

Определение

Выравниванием последовательности X к выравниванию w называется отображение $Q : (X, \bar{\Omega}) \rightarrow \times_{i=1}^{N+1} (\Sigma^{\max_{L \in L_i} (L)})$, 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 $(\Omega, \bar{\Omega})$ за пространство элементарных исходов Ω . Область значений выравнивания Q обозначим как $\bar{\Omega}$.

Определение

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

Введение

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

Определение

Выравниванием последовательности X к выравниванию w называется отображение $Q : (X, \bar{\Omega}) \rightarrow \times_{i=1}^{N+1} (\Sigma^{\max_{L \in L_i} (L)})$, 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 $(\Omega, \bar{\Omega})$ за пространство элементарных исходов Ω . Область значений выравнивания Q обозначим как $\bar{\Omega}$.

Определение

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

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

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

Оценивание значимости выравнивания

└ Введение

Введение

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

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

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

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

Оценивание значимости выравнивания

Введение

└ Введение

Так как последовательности описывают некоторые объекты, то введенную на предыдущем слайде оценку можно интерпретировать как оценку того, насколько сильно новая последовательность X близка к множеству последовательности, на которых было построено выравнивание w .

Встает вопрос того, как интерпретировать величину этой оценки:

- достаточно ли высокая эта оценка, чтобы считать объект, описываемый последовательностью X , родственным к объектам, описываемым выравниванием w , или шум, т.е. случайная последовательность, мог получить такую оценки.
- достаточно ли низкая эта оценка, чтобы считать объект описываемый последовательностью X , не родственным к объектам, описываемым выравниванием w , или сигнал, т.е. последовательность, описывающая взаимосвязанный с множеством объект, мог получить такую оценку.