Слайд 1

Титульный лист

Слайд 2

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.

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.

На слайде 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 вычислительно неэффективно, но оказывается, что аппарат скрытых марковских моделей (СММ) позволяет эффективно решать эту задачу.

Слайд 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.

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

Далее будет описаны метод, который позволяет эффективно вычислять введенный термин.

Слайд 4

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

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

Определение 1. *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");*
- *$P(Y_n = y_n | X_1 = x_1, \ldots, X_n = x_n) = P(Y_n | X_n = x_n)$ 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**).*

Слайд 5

Примером простой CMM 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.

Слайд 6

Профильная CMM — это CMM 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 CMM начинает и

заканчивает работу и не испускает наблюдаемых символов, как показано на слайде, 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. *Последовательность* D, связанная с путем π — последовательность наблюдаемых символов, которая была получена в результате прохода профильной СММ пути π.

Слайд 7

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.

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

$$s_{max}(D) = \underset{\pi \in \pi_D}{max}(s(\pi)), \tag{1}$$

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:

$$s_{fw}(D) = \sum_{\pi \in \pi_D} s(\pi). \tag{2}$$

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):

$$Z(D, T) = \sum_{\pi \in \pi_D} s(\pi)^{\frac{1}{T}}. \tag{3}$$

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

Слайд 8

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 $Pr(d|B)$, где d отражает возможный наблюдаемый символ:

$$Pr(D|B) = \prod_{i=1}^{L} Pr(d_i|B), \tag{4}$$

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

Слайд 9

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

$$fpr(s_0) = \sum_{D \in D_L} Pr(D|B)\Theta(s(D) \geq s_0), \tag{5}$$

где $Pr(D|B)$ — условная вероятность последовательности D, описываемая фоновой моделью, $s(D)$ — вероятность последовательности D, считаемая профильной СММ, 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 $s(D) = s_{fw}(D)$.

Слайд 10

Так как вычисление $fpr(s_0)$ через формулу 5 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 $P(D|T)$ — это условная вероятность строки D относительно некоторой модели строк длины L параметризованной значением T. Тогда можно переписать $fpr(s_0)$:

$$fpr(s_0) = \sum_{D \in D_L} Pr(D|T) f(D, s_0), \tag{6}$$

где

$$f(D, s_0) = \frac{Pr(D|B)\Theta(s(D) \geq s_0)}{Pr(D|T)}. \tag{7}$$

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 $fpr(s_0)$.

Слайд 11

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

$$Pr(D|T) = \frac{P(D|B)Z(D,T)}{Z(T)}, \tag{8}$$

где

$$Z(T) = \sum_{D \in D_L} Pr(D|B)Z(D,T). \tag{9}$$

Подставив определение $Pr(D,T)$ в уравнение 7 получим

$$f(D, s_0) = \frac{Z(T)\Theta(s(D) \geq s_0)}{Z(D,T)}. \tag{10}$$

В итоге мы хотим смоделировать последовательности в соответствии с распределением $Pr(D|T)$, вычислить $f(D, s_0)$ для каждой последовательности и использовать среднее этих значений как оценку $fpr(s_0)$.

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

Слайд 11

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, что вычислительно неосуществимо.