Потенциальные функции на множестве векторных последовательностей разной длины

Сулимова В. В., Моттль В. В., Мучник И. Б.
sulimova@tula.net, vmottl@yandex.ru, muchnik@dimacs.rutgers.edu
Тула, ТулГУ; Москва, ВЦ РАН; США, Rutgers University

Последовательности разной длины $\omega = (\alpha_k,\, k = 1, \ldots, N_\omega)$ являются типовыми объектами задач анализа данных. Примитивы $\alpha_k \in A$, составляющие последовательности, могут быть действительными числами, векторами или символами некоторого конечного алфавита. В первых двух случаях последовательности являются скалярными или векторными сигналами, а в последнем случае принято говорить о символьных последовательностях.

Широко известны такие задачи анализа последовательностей разной длины, как задача идентификации личности по динамике подписи, распознавание речевых команд и слитного текста, прогнозирование биологических свойств белков на основе анализа составляющих их последовательностей аминокислотных остатков.

Удобным инструментом решения таких задач являются методы, основанные на понятии потенциальной функции [1, 2] — действительной функции двух аргументов, матрица значений которой для любой конечной совокупности объектов является неотрицательно определенной. Потенциальная функция $K(\omega', \omega'')$, определенная на множестве произвольных объектов $\omega \in \Omega$, погружает исходное множество в некоторое линейное пространство $\tilde{\Omega} \supseteq \Omega$, в котором она является скалярным произведением. Такое погружение позволяет применять для объектов произвольной природы практически любые классические методы анализа данных, разработанные для линейных пространств, минуя промежуточный этап выбора вектора числовых признаков $\mathbf{x}(\omega) \in R^n$, определяющего скалярное произведение $K(\omega', \omega'') = \mathbf{x}^{\mathrm{T}}(\omega')\,\mathbf{x}(\omega'')$. Для последовательностей разной длины поиск числовых признаков особенно проблематичен.

Можно предложить разные способы введения потенциальной функции на множестве последовательностей разной длины. Мы сначала изложим одну достаточно общую математическую структуру потенциальной функции, адекватную потребностям многих практических задач, а затем рассмотрим частный случай этой структуры, приводящий к простому алгоритму вычисления потенциальной функции для любых двух заданных последовательностей.

Потенциальная функция на множестве примитивов

Будем полагать, что множество примитивов A является линейным пространством со скалярным произведением (потенциальной функцией)

$\mu(\alpha', \alpha'')$. Тогда величина $\sqrt{\mu(\alpha, \alpha)}$ играет роль нормы в этом линейном пространстве.

Интерпретация множества примитивов как линейного пространства представляется вполне естественной для сигналов, изначально являющихся последовательностями действительных чисел либо векторов. Можно показать, что такая интерпретация остается естественной и для аминокислотных последовательностей белков, представляющих собой символьные последовательности над алфавитом двадцати существующих в природе аминокислот, если используется общепринятый в молекулярной биологии способ измерения сходства всяких двух аминокислот как вероятности их происхождения в процессе эволюции от одной и той же неизвестной аминокислоты [3].

Множество выравниваний двух последовательностей

Если бы все последовательности имели одинаковую длину $\Omega = \{\omega = (\alpha_k,\, k = 1, \ldots, N = \text{const})\}$, то, например, произведение $K(\omega', \omega'') = \prod_{k=1}^{N} \mu(\alpha'_k, \alpha''_k)$ обладало бы всеми свойствами потенциальной функции на множестве Ω, поскольку известно, что произведение любого числа потенциальных функций также является потенциальной функцией [4]. Однако мы имеем дело с множеством последовательностей разной длины $\Omega = \{\omega = (\alpha_k,\, k = 1, \ldots, N_\omega)\}$, и применение такого способа возможно лишь после выравнивания длин сравниваемых последовательностей.

Под выравниванием w двух последовательностей $\omega' = (\alpha'_k,\, k = 1, \ldots, N')$ и $\omega'' = (\alpha''_k,\, k = 1, \ldots, N'')$, $\alpha'_k, \alpha''_k \in A$, понимают приведение их к одинаковой длине за счет добавления «пустых» выравнивающих элементов в некоторые позиции каждой из последовательностей с последующей перенумерацией элементов: $\bar{\omega}'_w = (\bar{\alpha}'_{w,j}\,,\, j = 1, \ldots, |w|)$ и $\bar{\omega}''_w = (\bar{\alpha}''_{w,j}\,,\, j = 1, \ldots, |w|)$, где $|w| \geqslant \max\{N', N''\}$ — общая длина выровненных последовательностей. В качестве выравнивающего элемента мы произвольно выберем некоторый элемент исходного линейного пространства примитивов $\alpha^0 \in A$, имеющий единичную норму $\mu(\alpha^0, \alpha^0) = 1$.

Множество всех выравниваний упорядоченной пары последовательностей $\langle \omega', \omega'' \rangle$ длин N' и N'' будем обозначать $\mathcal{W}_{N'N''}$. Любое выравнивание $w \in \mathcal{W}_{N'N''}$ может быть представлено в виде пути на графе с горизонтальными, диагональными и вертикальными ребрами, ориентированными слева направо и сверху вниз, как показано на Рис. 1. Горизонтальное направление на таком графе будем связывать с первой последовательностью $\omega' = (\alpha'_k,\, k = 1, \ldots, N')$, а вертикальное направление — со второй последовательностью $\omega'' = (\alpha''_k,\, k = 1, \ldots, N'')$.

Будем рассматривать всякое выравнивание $w \in \mathcal{W}_{N'N''}$ как последовательность значений переменной из трехэлементного алфавита:

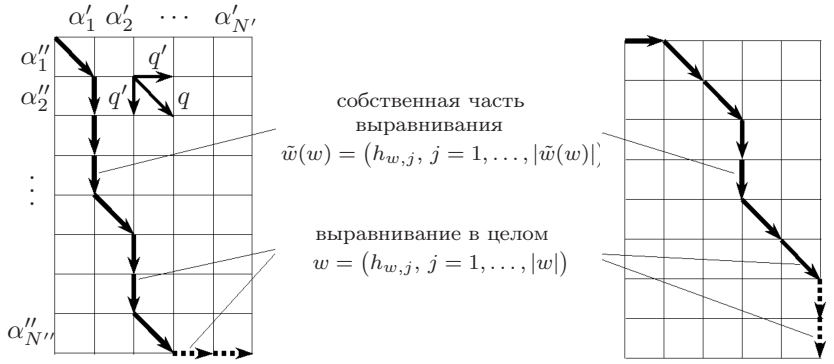

Рис. 1. Два разных выравнивания пары последовательностей.

$w = (h_k,\, k = 1, \ldots, |w|)$, $h_k \in \{h, h', h''\}$. Значение $h_k = h'$ означает продвижение на один шаг в горизонтальном направлении, т. е. вставку одного «пустого» выравнивающего элемента в первую последовательность, значение $h_k = h$ интерпретируется как диагональное продвижение, соответствующее отсутствию вставки, а $h_k = h''$ обозначает шаг в вертикальном направлении, вставляющий «пустой» элемент во вторую последовательность. Симметричный аналог всякого выравнивания w, получаемый заменой каждого шага $h_k = h'$ на $h_k = h''$ и наоборот, будем обозначать символом w^{T}, так что $\mathcal{W}_{N''N'} = \bigl\{w^{\mathrm{T}}\colon\, w \in \mathcal{W}_{N'N''}\bigr\}$.

Система весов на множестве выравниваний и структура потенциальной функции

С выравниванием w двух последовательностей ω' и ω'' свяжем действительную величину

$$K(\omega', \omega'' \,|\, w) = K(\omega'', \omega' \,|\, w^{\mathrm{T}}) = \prod_{j=1}^{|w|} \mu(\bar{\alpha}'_{w,j}, \bar{\alpha}''_{w,j}), \tag{1}$$

понимаемую как мера условного сходства двух последовательностей, зависящая от выбора выравнивания w.

Далее, выберем систему неотрицательных весов парных выравниваний $p(w) \geqslant 1$, общую для всех значений пар длин последовательностей N' и N'', и выражающую априорные предпочтения на множестве разных выравниваний одной и той же пары.

Традиционный способ измерения сходства двух последовательностей основан на поиске выравнивания, максимизирующего их условное сходство с учетом веса $K(\omega', \omega'') = \max\limits_{w \in \mathcal{W}_{N'N''}} p(w) K(\omega', \omega'' \,|\, w)$ [5], однако такая мера сходства не будет обладать свойствами потенциальной функции.

В данной работе вместо операции максимизации мы используем линейную комбинацию значений условного сходства двух последовательностей по всем выравниваниям с учетом их весов:

$$K(\omega',\omega'') = \sum_{w\in\mathcal{W}_{N'N''}} p(w)K(\omega',\omega''\,|\,w). \qquad (2)$$

Пусть w — некоторое выравнивание последовательностей ω' и ω'', имеющих длины N' и N''. Начальную часть выравнивания w до первого касания правой или нижней границы области $\mathcal{W}_{N'N''}$ (Рис. 1), будем называть его собственной частью и обозначать символом $\tilde{w}(w)$.

Дополним последовательности ω' и ω'' длин N' и N'' выравнивающими элементами $\alpha^0 \in A$ справа до некоторой длины N, и будем называть полученные последовательности $\bar{\omega}' = (\alpha'_{k'},\, k'{=}1,\ldots,N,\, \alpha'_{k'}{=}\alpha^0,\, k' > N')$ и $\bar{\omega}'' = (\alpha''_{k''},\, k''{=}1,\ldots,N,\, \alpha''_{k''}{=}\alpha^0,\, k'' > N'')$ расширенными. Все выравнивания расширенных последовательностей образуют множество $\mathcal{W}_{NN}$. Два выравнивания $w \in \mathcal{W}_{N'N''}$ и $\bar{w} \in \mathcal{W}_{NN}$ будем называть эквивалентными и обозначать как $w \sim \bar{w}$, если собственная часть выравнивания w является начальной частью выравнивания $\bar{w}$.

Система весов $p(w)$ называется согласованной, если, во-первых, веса симметричных выравниваний равны $p(w) = p(w^{\mathrm{T}})$, и, во-вторых, для любых N', N'' и N, таких, что $N \geqslant N'$ и $N \geqslant N''$, выполняется условие $p(w) = \sum_{\bar{w}\in\mathcal{W}_{NN},\, w\sim\bar{w}} p(\bar{w})$, т. е. вес выравнивания w исходных последовательностей равен сумме весов эквивалентных ему расширенных последовательностей.

Теорема 1. *Для того, чтобы линейная комбинация $K(\omega',\omega'')$ (2) условных мер сходства $K(\omega',\omega''\,|\,w)$ (1) обладала свойствами потенциальной функции на множестве последовательностей над линейным пространством примитивов $\Omega = \{\omega = (\alpha_k, k = 1,\ldots,N_\omega),\ \alpha_k{\in}A\}$, достаточно, чтобы выравнивающий элемент удовлетворял условию $\mu(\alpha^0,\alpha^0) = 1$ и система весов $p(w)$ была согласованной.*

Тот факт, что некоторая двухместная функция $K(\omega',\omega'')$ (2) формально обладает свойствами потенциальной функции на множестве последовательностей разной длины, еще не гарантирует ее практическую полезность. Важно удачно выбрать исходную потенциальную функцию $\mu(\alpha',\alpha'')$ на множестве примитивов $\alpha \in A$, выравнивающий элемент $\alpha^0 \in A$, а также систему весов выравниваний $p(w)$.

Радиальная потенциальная функция на множестве примитивов и мультипликативные веса выравниваний

Пусть в линейном пространстве примитивов с нулевым элементом $\varnothing \in A$ определена евклидова метрика, например, с помощью некото-

рой исходной потенциальной функцией $\rho(\alpha', \alpha'') = \big[\varkappa(\alpha', \emptyset) + \varkappa(\alpha'', \emptyset) - 2\varkappa(\alpha', \alpha'')\big]$. Известно [1], что в этом случае двухместная функция

$$\mu(\alpha', \alpha'') = \exp\big[-\beta\rho^2(\alpha', \alpha'')\big] \tag{3}$$

обладает свойствами потенциальной функции при любом значении параметра $\beta{>}0$, преобразуя линейное пространство A в некоторое другое линейное пространство со скалярным произведением $\mu(\alpha', \alpha'')$.

Потенциальную функцию (3), по своему смыслу являющуюся мерой сходства примитивов относительно исходной метрики $\rho(\alpha', \alpha'')$, принято называть радиальной.

Выбор нулевого элемента исходного линейного пространства в качестве выравнивающего элемента $\alpha^0 = \emptyset \in A$ удовлетворяет условию $\mu(\alpha^0, \alpha^0) = 1$ в Теореме 1.

С каждым из трех значений переменной h, h' и h'' свяжем неотрицательные числа $q(h) = q$ и $q(h') = q(h'') = q'$, $q + 2q' = 1$. Значение $q > 1/3$ задает предпочтительность отсутствия вставок и удалений элементов на каждом элементарном шаге сравнения последовательностей, Рис. 1.

Пусть $w = (h_{w,j}, j = 1, \ldots, |w|)$ — произвольное выравнивание, $\tilde{w}(w)$ — его собственная часть. Вес выравнивания определим как произведение

$$p(w) = \prod_{j=1}^{|\tilde{w}(w)|} q(h_{w,j}). \tag{4}$$

Теорема 2. *Система весов выравниваний (4) является согласованной.*

Таким образом, радиальная потенциальная функция на множестве примитивов и мультипликативная система весов выравниваний удовлетворяют всем требованиям Теоремы 1 и определяют потенциальную функцию на множестве последовательностей разной длины (2), явным образом выражающую степень их попарного сходства. Алгоритм вычисления такой потенциальной функции имеет сложность, пропорциональную произведению длин сравниваемых последовательностей.

Работа выполнена при поддержке РФФИ, проекты № 05-01-00679, № 06-01-08042 и № 06-01-00412, а также INTAS, проекты № 04-77-7347 и № 06-1000014-6563;

Литература

[1] *Айзерман М. А., Браверманн Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. — 384 с.

[2] *Vapnik V.* Statistical Learning Theory. — New York: John-Wiley & Sons, Inc., 1998. — 732 p.

[3] *Dayhoff M. O., Schwartz R. M., Orcutt B. C.* A model for evolutionary change in proteins. — Atlas for Protein Sequence and Structure (M. O. Dayhoff, ed.). — 1978. — Vol. 5. — Pp. 345–352.

[4] *Haussler D.* Convolution kernels on discrete structures. — Technical Report UCSC-CLR-99-10, University of California at Santa Cruz, 1999.

[5] *Dubin R., Eddy S., Krogh A., Mitchison G.* Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids. — Cambridge University Press, 1998. — 356 p.