УДК 519.6

УВЕЛИЧЕНИЕ РАЗМЕРНОСТИ В МЕТОДЕ ДОКИНГА НА ОСНОВЕ ТЕНЗОРНЫХ ПОЕЗДОВ

Д. А. Желтков[1], Е. Е. Тыртышников[2]

Предложена модификация метода докинга на основе тензорных поездов с использованием идеи искусственного увеличения размерности при поиске положения молекулы-лиганда, минимизирующего энергию системы белок–лиганд. Проведено сравнение с программой TTDock (докинг на основе тензорных поездов), не использующей искусственное увеличение размерности. Результаты тестирования показывают, что при той же надежности предложенная модификация в 5–10 раз вычислительно менее затратна, чем TTDock.

Ключевые слова: тензорный поезд, искусственное увеличение размерности, крестовый метод, глобальная оптимизация, докинг, компьютерный дизайн лекарств.

1. Введение. Многомерные массивы (тензоры) можно представлять в формате тензорного поезда (TT, от англ. Tensor Train) [1]. Необходимые определения можно также найти в [2]. Если TT-ранги d-мерного тензора $A \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ равны $r_0, \ldots, r_d$, то для его записи в TT-формате требуется $O(dnr^2)$ параметров (ячеек памяти), где

$$n = \max(n_1, \ldots, n_d), \quad r = \max(r_0, \ldots, r_d).$$

Таким образом, при относительно малом значении r используется $O(\log N)$ параметров, где N — общее число элементов, т.е. память для хранения тензора в TT-формате зависит от общего числа его элементов логарифмически.

Отметим, что основные операции в TT-формате выполняются за $O(dnr^3)$ действий. Для нас особенно важен метод TT-CROSS [3], позволяющий получить TT-приближение тензора по небольшому числу элементов аппроксимируемого тензора. Таких элементов нужно всего $O(dnr^2)$. Стоимость вычисления складывается из $O(dnr^3)$ арифметических операций и $O(dnr^2)$ операций поиска соответствующих элементов тензора. Если каждый элемент представляет собой значение некоторой функции в соответствующем узле массива, то требуется $O(dnr^2)$ раз вычислять значения указанной функции.

Идея искусственного увеличения числа измерений хорошо известна, например, в программировании: одномерный массив (вектор) часто рассматривается как двумерный массив (матрица), а иногда и как массив с числом измерений $d \geqslant 3$ (в частности, при развертке циклов). Введение виртуальных (искусственных) измерений для тензоров было предложено в [4]; в этой же работе впервые даны оценки величин, теперь называемых QTT-рангами тензора. Логарифмичность числа операций и хранимых элементов при использовании TT-формата делает идею искусственного увеличения размерности особенно привлекательной [5, 6].

Для простоты изложения будем считать, что $n_1 = \ldots = n_d = n$. Пусть $n = 2^m$. Преобразуем заданный d-мерный тензор $A \in \mathbb{R}^{n \times \cdots \times n}$ в dm-мерный тензор $B \in \mathbb{R}^{2 \times \cdots \times 2}$. Это можно сделать многими способами, например так:

$$B(i_{11}, \ldots, i_{m1}, i_{12}, \ldots, i_{m2}, \ldots, i_{1d}, \ldots, i_{md}) = A(j_1, j_2, \ldots, j_d),$$

где $j_k = i_{1k} + 2i_{2k} + \cdots + 2^{m-1} i_{mk}$.

Если тензор A в TT-формате занимает $O(dnr^2)$ ячеек памяти и операции с ним требуют $O(dnr^3)$ арифметических действий, то тензор B в том же формате занимает $O(dmR^2) = O(dR^2 \log_2 n)$ ячеек памяти и операции с ним требуют $O(dmR^3) = O(dR^3 \log_2 n)$ действий. Как правило, $R \geqslant r$; в ряде случаев такое преобразование может не привести к уменьшению объема хранимых данных и/или ускорению выполнения операций с тензором. Однако часто значение R близко к r, а в таких случаях мы получаем существенный выигрыш как по памяти, так и по числу операций.

[1] Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, Ленинские горы, д. 1, стр. 52, 119992, Москва; аспирант, e-mail: dmitry.zheltkov@gmail.com

[2] Институт вычислительной математики РАН, ул. Губкина, д. 8, 119333, Москва; директор, e-mail: eugene.tyrtyshnikov@gmail.com

В настоящей статье идея искусственного увеличения размерности применена к методу докинга на основе тензорных поездов [2]. Программа докинга TTDock [2] преобразована в программу докинга QTTDock. Сравнение этих программ на одних и тех же данных показало, что при той же надежности программа QTTDock превосходит TTDock в 5–10 раз по быстродействию.

2. Программа QTTDock. Основная задача докинга — поиск глобального минимума и близких к нему по значению локальных минимумов энергии системы белок–лиганд. Для ее решения в [2] предложен новый подход, в котором задача оптимизации по степеням свободы лиганда решается на основе тензорных поездов и метода TT-CROSS [3]. В настоящей работе мы предлагаем модификацию метода TTDock с использованием искусственного увеличения размерности.

Пусть лиганд имеет d степеней свободы. Для простоты изложения будем считать, что область изменения каждой степени свободы представляет собой отрезок $[0, 1]$. На каждом таком отрезке введем равномерную сетку с $n = 2^m$ узлами и шагом $h = 1/(2^m - 1)$. В соответствии с методом TTDock заменим исходную оптимизируемую функцию вычисления энергии системы белок–лиганд $e(\bar{v})$ на функционал $g_\alpha(\bar{(v)}) = f(e(\bar{v} - \alpha))$, где $f(x)$ — непрерывное монотонно убывающее отображение интервала $(-\infty, +\infty)$ в подобласть полуинтервала $[0, +\infty)$ и α — текущее приближение к глобальному минимуму. Однако вместо формирования тензора $A^\alpha \in \mathbb{R}^{2^m \times \cdots \times 2^m}$ размерности d сформируем тензор $B^\alpha \in \mathbb{R}^{2 \times \cdots \times 2}$ размерности $D = dm$:

$$B^\alpha(i_{11}, \ldots, i_{m1}, i_{12}, \ldots, i_{m2}, \ldots, i_{1d}, \ldots, i_{md}) = g_\alpha(j_1 h, j_2 h, \ldots, j_d h),$$

где $j_k = i_{1k} + 2i_{2k} + \cdots + 2^{m-1} i_{mk}$.

Далее выполним с тензором B^α операции, аналогичные операциям с тензором A^α в методе TTDock: последовательно от $k = 1$ до $D - 1$ (шаги слева направо) и от $k = D - 1$ до 1 (шаги справа налево) приближаем подматрицы $B^\alpha_k(I_k, J_k)$ матриц разверток

$$B^\alpha_k(i_1 \ldots i_k, i_{k+1} \ldots i_D) = B^\alpha(i_1, \ldots, i_D)$$

крестовым методом интерполяции матриц [8–11] с рангом не выше некоторого $r_{\max}$, локально оптимизируем полученные узлы интерполяции и добавляем их в множество локальных минимумов M. С помощью точек интерполяции и точек, полученных из них локальной оптимизацией, переопределяем множество строк I_k и множество столбцов J_k. Используя I_k и J_k, формируем I_{k+1} и J_{k+1} при шагах слева направо или I_{k-1} и J_{k-1} при шагах справа налево. Подробно метод описан в [2]. Метод выполняется итерационно, число итераций и ограничение ранга $r_{\max}$ подбираются в ходе работы.

Отметим, что в программе QTTDock локальные оптимизации производятся чаще: их $O(Dr_{\max}) = O(dmr_{\max}) = O(dr_{\max} \log_2 n)$, а в TTDock их $O(dr_{\max})$. Кроме того, для достижения той же надежности, что и в методе TTDock, теперь можно было бы ожидать большего ранга $r_{\max}$ или большего числа итераций. Однако практические эксперименты показывают, что ранг и число итераций не растут, а увеличение числа локальных оптимизаций влияет не столь существенно, как уменьшение числа вычисленных значений функции.

Сложность метода составляет:

$O(Dr^3_{\max}) = O(dmr^3_{\max}) = O(dr^3_{\max} \log_2 n)$ операций,

$O(Dr^2_{\max}) = O(dmr^2_{\max}) = O(dr^2_{\max} \log_2 n)$ вычислений функции и

$O(Dr^3_{\max}) = O(dmr^3_{\max}) = O(dr^3_{\max} \log_2 n)$ локальных оптимизаций.

Программа QTTDock реализована на языке C++. В качестве метода локальной оптимизации используется реализация симплекс-метода библиотеки GSL (GNU Scientific Library) [12].

3. Сравнение программ TTDock и QTTDock. Тестирование проводилось с использованием целевой функции и пар белок–лиганд, подробно описанных в [2]. Строились сетки размера $n = 2^7 = 128$ вдоль каждого направления. Для каждой пары белок–лиганд произведено по 1000 экспериментов каждой из программ. Надежностью назовем долю экспериментов, в которых рассматриваемым методом найден глобальный минимум, сложностью — среднее количество вычислений энергии системы белок–лиганд. Результаты сравнения приведены в таблице.

Как показывают результаты тестирования, программа QTTDock в 5–10 раз быстрее программы TTDock при том же уровне надежности.

Заключение. В настоящей статье применена идея искусственного увеличения размерности к методу докинга на основе тензорных поездов. Полученная программа QTTDock имеет тот же уровень надежности, что и программа TTDock, однако требует значительно, в 5–10 раз меньше вычислений энергии системы белок–лиганд. Такое ускорение позволяет реализовать докинг без использования предварительно

Сложность и надежность программ TTDock/QTTDock

Белок–лиганд	Сложность	Надежность
chk1_8	3.0×10^6 / 5.4×10^5	1 / 1
urokinase_7	2.3×10^7 / 4.9×10^6	1 / 1
erk2_000124	7.5×10^7 / 7.5×10^6	0.98 / 0.97

рассчитанной сетки потенциалов, в том числе докинг для подвижных белков-мишеней, что в конечном итоге должно привести к увеличению эффективности применения докинга для разработки новых лекарств.

Работа выполнена при финансовой поддержке Минобрнауки России по государственному контракту от 11.03.2013 г. № 14.514.11.4070 в рамках ФЦП “Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 годы”.

СПИСОК ЛИТЕРАТУРЫ

1. *Oseledets I.V., Tyrtyshnikov E.E.* Breaking the curse of dimensionality, or how to use SVD in many dimensions // SIAM J. Sci. Comput. 2009. **31**, N 5. 3744–3759.
2. *Желтков Д.А., Офёркин И.В., Каткова Е.В., Сулимов А.В., Сулимов В.Б., Тыртышников Е.Е.* TTDock: метод докинга на основе тензорных поездов // Вычислительные методы и программирование. 2013. **14**. 279–291.
3. *Oseledets I.V., Tyrtyshnikov E.E.* TT-cross approximation for multidimensional arrays // Linear Algebra Appl. 2010. **432**, N 1. 70–88.
4. *Тыртышников Е.Е.* Тензорные аппроксимации матриц, порожденных асимптотически гладкими функциями // Математический сборник. 2003. **194**, № 6. 146–160.
5. *Oseledets I.V.* Approximation of $2^d \times 2^d$ matrices using tensor decomposition // SIAM J. Matrix Anal. Appl. 2010. **31**, N 4. 2130–2145.
6. *Khoromkij B.N., Oseledets I.V.* QTT approximation of elliptic operators in higher dimensions // Rus. J. Numer. Anal. Math. Model. 2011. **26**, N 3. 303–322.
7. *Oseledets I.V.* Tensor-train decomposition // SIAM J. Sci. Comput. 2011. **33**, N 5. 2295–2317.
8. *Goreinov S.A., Tyrtyshnikov E.E., Zamarashkin N.L.* A theory of pseudo-skeleton approximations // Linear Algebra Appl. 1997. **261**, N 1–3. 1–21.
9. *Tyrtyshnikov E.E.* Incomplete cross approximation in the mosaic-skeleton method // Computing. 2000. **64**, N 4. 367–380.
10. *Goreinov S.A., Tyrtyshnikov E.E.* The maximal-volume concept in approximation by low-rank matrices // Contemporary Mathematics. 2001. **208**. 47–51.
11. *Goreinov S.A., Oseledets I.V., Savostyanov D.V., Tyrtyshnikov E.E., Zamarashkin N.L.* How to find a good submatrix // Matrix Methods: Theory, Algorithms, Applications / Edited by V. Olshevsky and E. Tyrtyshnikov. Hackensack: World Scientific, 2010. 247–256.
12. GNU Scientific Library (http://www.gnu.org/software/gsl).

Поступила в редакцию
15.04.2013