Математические Методы Прогнозирования

Кафедра Интеллектуальных Систем

Отчёты о лабораторной работе №2

Оглавление

Глава 1
Muzalevskiy

1.1. Введение

Тензор - это многомерный массив. Следовательно, практически все геометрические структуры данных, с которыми мы работаем, являются тензорами. До размерности $d = 2$ эти тензоры имеют специфические названия: скаляр, вектор, матрица. Данные структуры представлены на рис.1.1

Декомпозиция называют процесс разбиения на составные элементы. По сути, это означает факторизацию тензора размерности d. Данный процесс заключается в нахождении оптимального разбиения общей структуры на элементы. В общем случае декомпозиция мотивируется необходимостью получить более простую совокупность составляющих элементов, которые могут наилучшим образом представить данную систему (или данные) [**decomposition**].

Прежде чем описать трехмерное разложение, опишем для начала принципы двумерного разложения (т.е. разложения матрицы). Подходы к двумерному разложению хорошо известны и включают анализ главных компонент (PCA), анализ независимых компонент (ICA), неотрицательную матричную факторизацию (NMF) и анализ разреженных компонент (SCA). Эти методы стали стандартными инструментами, например, для разделения источников (BSS), извлечения признаков или классификации [**matrixdecomp**].

$$X \approx M = \sum_{r=1}^{R} a_r \cdot b_r^T = a_r \circ b_r = A \cdot B^T X \in \mathbb{R}^{I \times J}, \mathbf{a} \in \mathbb{R}^I, \mathbf{b} \in \mathbb{R}^J$$

Где R - новая (уменьшенная) размерность наших данных, часто называемая рангом. Эта операция представляет собой простое суммирование внешних про-

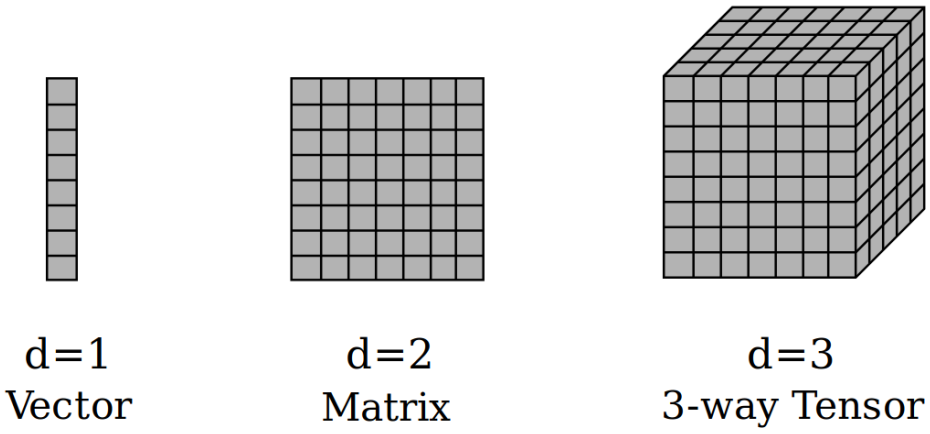

Рис. 1.1. "Структуры"

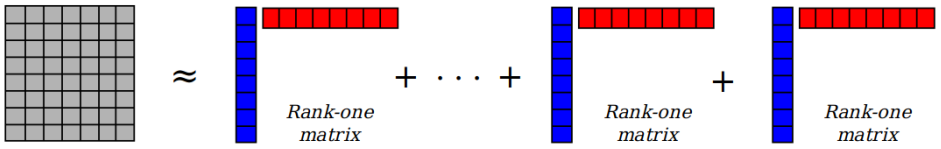

Рис. 1.2. "Факторизация"

изведений каждого столбца A и B, где индекс столбца задается r, как показано ниже на рис.1.2

Такая декомпозиция известна как факторизация. Приведенная выше формулировка страдает от проблемы, называемой проблемой вращения(rotation problem). То есть, мы можем вставить любую несингулярную матрицу вращения, Z, в приведенную выше формулировку, и в итоге получим то же самое приближение X.

$$X \approx M = \sum_{r=1}^{R} a_r \circ z_r^T \circ z_r^{-1} \circ b_r^T = A \cdot Z^T \cdot Z^{-1} \cdot B^T$$

Следовательно, если приведенная выше формулировка не ограничена, то она приводит к бесконечному множеству комбинаций A и B. Стандартные методы факторизации матриц в линейной алгебре, такие как QR-факторизация, разложение по собственным значениям (EVD) и разложение по сингулярным значениям (SVD), являются лишь частными случаями вышеприведенной формулировки и обязаны своей единственностью ограничениям, таким как треугольность и ортогональность.

Трехмерная декомпозиция - это просто расширение двумерной декомпозиции. Однако, хотя в двумерном случае для получения уникального решения на задачу должны быть наложены явные ограничения, высокая размерность тензорного формата имеет свои преимущества - это возможность получения компактных представлений, уникальность разложения, гибкость в выборе ограничений и общность компонентов, которые могут быть идентифицированы.

$$X \approx M = \sum_{r=1}^{R} a_r \circ b_r \circ c_r X \in \mathbb{R}^{I \times J \times K}, \mathbf{a} \in \mathbb{R}^I, \mathbf{b} \in \mathbb{R}^J, \mathbf{c} \in \mathbb{R}^K$$

В результате такого разложения мы получим три матрицы A с размерностью (IR), B с размерностью (JR) и C с размерностью (KR). Эта операция является простым суммированием внешнего произведения каждого столбца A, Bи C, где индекс столбца задан r, как показано на рис.1.3

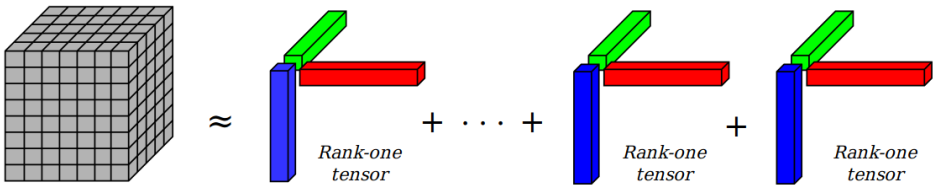

Рис. 1.3. "Разложение"

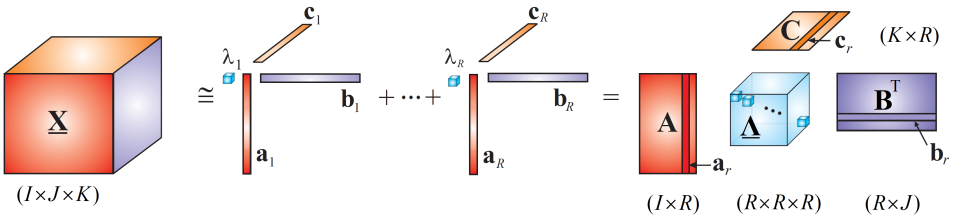

Рис. 1.4. "Canonical Polyadic Decomposition"

1.2. Постановка задачи

В данной работе, мы рассмотрим основные алгоритмы для тензорной декомпозиции, а также их реализации в библиотеке *hottbox* [**hottbox**]. В секции "Вычислительный эксперимент и анализ ошибки мы проведем сравнение ошибки данных методов, используя отношения норм Фробениуса[**matrixnorm**]

1.3. Модели

В рамках данной секции, мы рассмотрим две модели, представленные в пакете *hottbox* - это модель Canonical Polyadic Decomposition (CPD) [**CPD**] и модель Higher Order Singular Value Decomposition ($HOSVD$) [**HOSVD**].

Canonical Polyadic Decomposition (CPD) (также называемое $PARAFAC$ или $CANDECOMP$) - это алгоритм, который факторизует тензор 3-го порядка $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ в линейную комбинацию членов $\underline{\mathbf{X}}_r = \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$. Другими словами, тензор раскладывается следующим образом:

$$\begin{aligned} \underline{\mathbf{X}} &\simeq \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \\ &= \underline{\mathbf{\Lambda}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} \\ &= [\underline{\mathbf{\Lambda}}; \mathbf{A}, \mathbf{B}, \mathbf{C}] \end{aligned}$$

В *hottbox* алгоритм CPD (с использованием метода Alternating Least Squares) реализован классом CPD(). Он выводит экземпляр класса $TensorCPD()$ после каждого вызова метода $decompose()$. Этот метод принимает объект класса Tensor

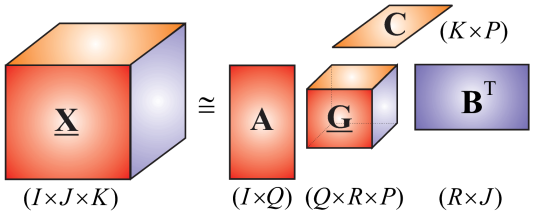

Рис. 1.5. "Tucker Decomposition"

и желаемое значение ранга Крускала, передаваемое в виде кортежа длины 1. Ранг Крускала передается в виде кортежа, чтобы сохранить одинаковый формат с другими алгоритмами тензорных разложений.

Перед тем, как перейти к рассмотрению алгоритма Higher Order Singular Value Decomposition ($HOSVD$), опишем принцип, на котором базируется данный алгоритм. Данный принцип носит название Разложение Такера (Tucker Decomposition).

Разложение Такера представляет заданный тензор $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ в виде тензора с плотным ядром $\underline{\mathbf{G}}$ и набором факторных матриц $\mathbf{A} \in \mathbb{R}^{I \times Q}, \mathbf{B} \in \mathbb{R}^{J \times R}$ и $\mathbf{C} \in \mathbb{R}^{K} \times P$ Другими словами, тензор $\underline{\mathbf{X}}$ может быть представлен в форме Такера следующим образом:

$$\begin{aligned}\underline{\mathbf{X}} &\simeq \sum_{q=1}^{Q}\sum_{r=1}^{R}\sum_{p=1}^{P} g_{qrp}\mathbf{a}_q \circ \mathbf{b}_r \circ \mathbf{c}_p \\ &= \underline{\mathbf{G}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} \\ &= [\mathbf{G}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\end{aligned}$$

Higher Order Singular Value Decomposition($HOSVD$) - частный случай разложения Таккера, в котором все матрицы факторов ограничены ортогональностью. Они вычисляются как усеченная версия сингулярных матриц всех возможных поворотов тензора.

$$\begin{aligned}\mathbf{X}_{(1)} &= \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T \quad \rightarrow \quad \mathbf{A} = \mathbf{U}_1[1:R_1] \\ \mathbf{X}_{(2)} &= \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^T \quad \rightarrow \quad \mathbf{B} = \mathbf{U}_2[1:R_2] \\ \mathbf{X}_{(3)} &= \mathbf{U}_3\mathbf{\Sigma}_3\mathbf{V}_3^T \quad \rightarrow \quad \mathbf{C} = \mathbf{U}_3[1:R_3]\end{aligned}$$

Причем, мы рассматриваем трехмерный тензор $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ как следующее представление(после применения разложения Такера):

$$\underline{\mathbf{X}} = \underline{\mathbf{G}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$$

Для тензора общего порядка - кортеж N-tuple $(R_1, \ldots, R_N)$ называется мультилинейным рангом и обеспечивает гибкость при сжатии и аппроксимации исходного тензора. После получения матриц коэффициентов, основной тензор вычисля-

ется как

$$\underline{\mathbf{G}} = \mathbf{X} \times_1 \mathbf{A}^T \times_2 \mathbf{B}^T \times_3 \mathbf{C}^T$$

1.4. Вычислительный эксперимент и анализ ошибки

Ссылка на код работы

Для расчета результатов работы алгоритмов, был сгенерирован синтетический датасет размерности 100x100x100 с добавлением шумовой компоненты.

Ошибка считалась с помощью встроенного метода "residual rel error который рассчитывает отношение норм Фробениуса двух тензоров (изначального и полученного в результате применения алгоритма).

В данной работе мы получили примерно одинаковые результаты работы алгоритмов CPD и $HOSVD$ - относительная ошибка аппроксимации составила 0.44.

Разумеется, остается большой простор для сравнения различных датасетов и применения алгоритмов для решения практических задач.