УДК 004.85
doi:10.18720/SPBPU/2/id23-510

Малич Виктория Олеговна [1],
студент магистратуры;
Нестеров Сергей Александрович [2],
доцент, канд. техн. наук, доцент

МЕТОДЫ СНИЖЕНИЯ РАЗМЕРНОСТИ ДЛЯ ЗАДАЧ АНАЛИЗА ДАННЫХ

[1, 2] Россия, Санкт-Петербург,
Санкт-Петербургский политехнический университет Петра Великого,
[1] malichvo@gmail.com, [2] nesterov@spbstu.ru

Аннотация. В статье рассматриваются методы снижения размерности данных. В качестве примера используется сингулярное разложение, в результате которого декомпозируется разрежённая матрица с оценками пользователей. Восстановленная матрица проверяется на точность с помощью регрессионных метрик. Полученная модель используется для вывода пользовательских рекомендаций.

Ключевые слова: снижение размерности, анализ данных, рекомендательная система, сингулярное разложение.

Viktoria O. Malich [1],
Master's Student;
Sergey A. Nesterov [2],
Candidate of Technical Sciences, Associate Professor

DIMENSIONAL REDUCTION METHODS FOR DATA ANALYSIS

[1, 2] Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russia,
[1] malichvo@gmail.com, [2] nesterov@spbstu.ru

Abstract. The paper discusses methods to reduce the dimensionality of the data. As an example, a singular decomposition is used which decomposes a sparse matrix with user ratings. The reconstructed matrix is tested for accuracy using regression metrics. The resulting model is used to derive user recommendations.

Keywords: dimensional reduction, data analysis, recommender system, singular value decomposition.

Введение

Многомерные данные можно преобразовать в вектор чисел. Например, матрицу пикселей изображения можно записать в виде вектора. Однако, такие векторы обычно достаточно длинные и содержащиеся в них признаки неинформативны.

Снижение размерности данных позволяет получить компактное множество информативных признаков, поскольку избыточные или неинформативные признаки приводят к понижению эффективности и точности модели.

Одним из примеров задачи анализа данных, где используется методы снижения размерности, является построение рекомендаций.

Рекомендательные системы прочно закрепились в различных сферах деятельности человека и улучшают пользовательский опыт. Персонализация пользовательского опыта напрямую влияет на доход компании, поскольку повышается вовлеченность клиентов. Рекомендательные системы предлагают товары в интернет-магазинах, составляют плейлисты или подборку научных статей, рекомендуют к просмотру видео или фильмы. В связи с большим количеством информации человек физически не имеет возможности ознакомиться со всем разнообразием, поэтому тема рекомендательных систем актуальна для принятия решения.

1. Методы снижения размерности

Снижение размерности данных достигается методами выбора признаков или извлечения признаков.

При выборе признаков остается некоторое подмножество исходного набора признаков, в котором отсутствуют избыточные и слабо информативные признаки, а также не создаются новые сложные признаки.

При извлечении признаков уменьшают размерность входных данных, составляя из исходных признаков новые, полностью описывающие пространство набора данных.

Один из методов, который относится к выбору признаков, называется метод опорных векторов [1]. Данный метод является итеративным. По исходному набору данных обучается классификатор, благодаря которому происходит дальнейшее ранжирование по весам. Оптимальным образом по полученным весам отсекается определенное количество признаков. Новое множество используется для повторного обучения классификатора. Алгоритм повторяется до тех пор, пока не будет выполнено условие по достижению требуемого количества признаков.

К методам извлечения признаков относится метод главных компонент [2]. Основная идея заключается в линейном ортогональном преобразовании, за счет чего данные из исходного признакового пространства отображаются в новое подпространство с меньшей размерностью. В качестве осей нового подпространства выступают первая и вторая главная компонента. Дисперсия данных каждой из осей максимизируется.

2. Типы рекомендательных систем

Выделяют четыре типа рекомендательных систем [3].

В методе контентных рекомендаций по базовым данным товара рекомендуется аналогичный товар.

Рекомендации на основе знаний строятся на более подробной информации о свойствах предмета, в отличие от контентных рекомендаций.

В коллаборативной фильтрации рекомендации строятся по известным предпочтениям группы пользователей.

Гибридные рекомендательные системы совмещают возможности базовых подходов.

3. Постановка задачи

Заданы два множества. Множество пользователей:

$$U = \{u_1, \ldots, u_n\}, \tag{1}$$

множество объектов, с которыми пользователи взаимодействовали:

$$I = \{i_1, \ldots, i_m\}. \tag{2}$$

Результат взаимодействия пользователя u и объекта i – r_{ui}.

Необходимо восстановить зависимость и вычислить

$$f(u,i) = \hat{r}_{ui} \approx r_{ui}\,. \tag{3}$$

Задача сводится к минимизации ошибки:

$$\|r_i - \hat{r}_i\|_2 \to \min\,. \tag{4}$$

4. Уменьшение размерности

Рассмотрим алгоритм, использующийся в коллаборативных рекомендательных системах, который основывается на разложении матриц, что уменьшает размерность исходного набора данных.

Поскольку используемый набор данных представлен в виде разрежённой матрицы, применяем сингулярное разложение (SVD) для сокращения признакового пространства и формирования нового, где эта разрежённость будет отсутствовать.

Матрица оценок $F \in R^{n\times m}$ раскладывается на произведение трех матриц:

$F = VDU^T =$

$$= \begin{pmatrix} | & | & | & | & | \\ \upsilon_1 & \ldots & \upsilon_k & \ldots & \upsilon_n \\ | & | & | & | & | \end{pmatrix} \begin{pmatrix} \sigma_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_k & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} | & | & | & | & | \\ u_1 & \ldots & u_k & \ldots & u_m \\ | & | & | & | & | \end{pmatrix}^T, \tag{5}$$

где $V \in R^{m\times n}$ и $U^T \in R^{n\times m}$ — ортогональные матрицы, столбцами которых являются левые и правые сингулярные векторы матрицы F;

$D = diag(\sigma_1, \sigma_2, ..., \sigma_k, 0, ..., 0) \in R^{n\times n}$ — диагональная матрица, где элементы $\sigma_1 \geq \sigma_2 \geq ..., \geq \sigma_k$ есть сингулярные числа матрицы F, $\sigma_1 = \|F\|$; при этом $0 < k < n$;

F — разрежённая user-item матрица;

$\hat{F}$ — восстановленная матрица, т. е. используя k сингулярных векторов, раскладываем матрицу F на компоненты V_k, D_k, U_k^T, результатом их перемножения является матрица $\hat{F}$:

$$\begin{pmatrix} | & | & | \\ \upsilon_1 & \cdots & \upsilon_k \\ | & | & | \end{pmatrix}\begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & 0 & \sigma_k \end{pmatrix}\begin{pmatrix} | & | & | \\ u_1 & \cdots & u_k \\ | & | & | \end{pmatrix}^T =$$

$$= V_k D_k U_k^T \approx \hat{F}, \# (6)$$

$V_k \in R^{m\times k}, D_k \in R^{k\times k}, U_k \in R^{k\times n}$.

Для хранения данных матрицы $F \in R^{n\times m}$ необходимо выделить память под nm элементов.

Для хранения данных матрицы $\hat{F} \in R^{n\times m}$ необходимо выделить память под $mk + k + kn = (m + n + 1)k$ элементов.

Таким образом, сингулярное разложение позволяет добиться уменьшения затрачиваемой памяти, а также увеличивает производительность.

5. Решение задачи

Набор данных, используемый для решения задачи, находится в открытом доступе на сайте kaggle.com [4] под названием «KinoPoisk movies and votes». Для получения уникальных идентификаторов пользователей и фильмов, а также оценок к фильмам используется файл movie_votes.csv, содержание файла изображено в таблице 1. Файл movie_info.csv содержит информацию о фильмах, необходим для вывода названия в рекомендательной системе.

Таблица 1

Содержание файла movie_votes.csv

№	user_id	movie_id	score	time
0	15647798	568289	9	1542847800
1	15647798	435	8	1542847860
…	…	…	…	
33219316	1510864	17176	6	1612831740

Создается матрица F размерности 13893×129570, в которую заносятся оценки каждого пользователя. В итоге получается разрежённая матрица, изображенная в таблице 2.

Таблица 2

Матрица оценок пользователей

0	0	…	0	0
6	8	…	0	0
…	…	…	…	…
0	0	…	0	0

Пусть количество сингулярных чисел k = 3681, что объясняет 90 % дисперсии (рис. 1). Тогда разряженная матрица F раскладывается на матрицы V_k размерности 13893×3681, D_k размерности 3681×3681 и U_k размерности 3681×129570.

Таким образом, для хранения матрицы оценок пользователей требуется выделять память под 13893×129570 элементов, а для хранения матриц $V_kD_kU_k$ 3681(13893+129570+1) элементов, что в 3,4 раза меньше исходной матрицы.

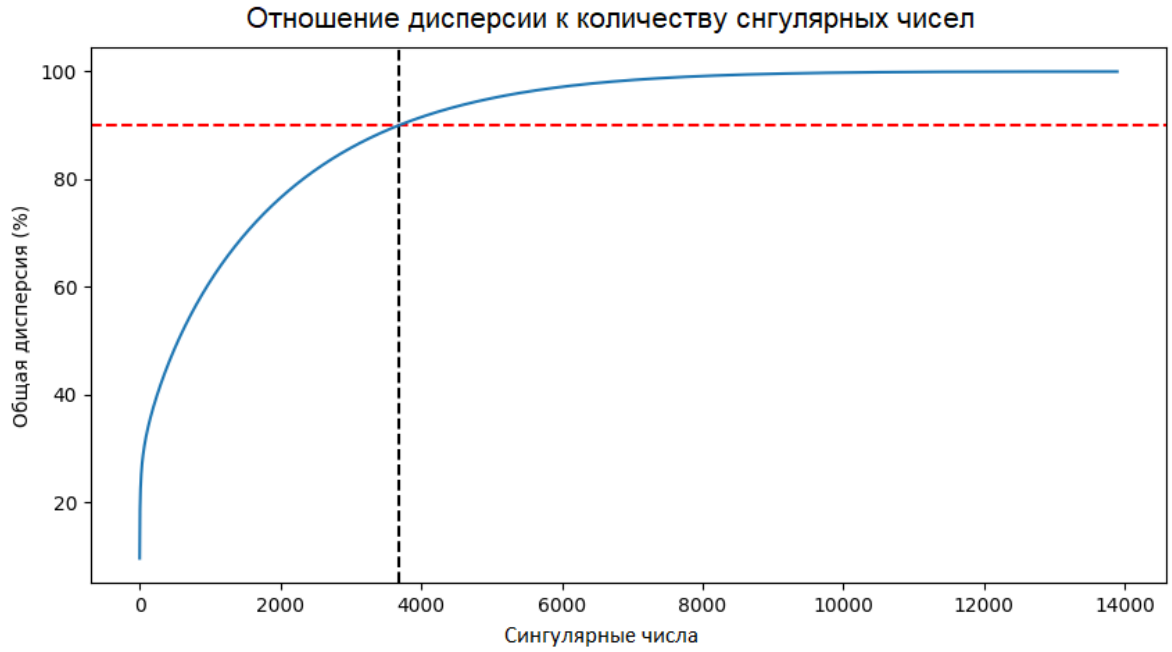

Рис. 1. Количество сингулярных чисел, объясняющие 90 % дисперсии

Перемножив матрицы $V_kD_kU_k$, новые значения восстановленной матрицы $\hat{F}$ можно трактовать как прогноз оценки пользователей. Для получения рекомендаций необходимо упорядочить список предсказанных оценок по убыванию.

```
Фильмы, рекомендуемые пользователю 15647798:
[0.2250633  0.19538594 0.19399499 0.18270293 0.18170438 0.17852709]
['Восстание планеты обезьян', 'Еще по одной', 'Человек-паук: Через вселенные',
 'Не говори никому', 'Величайший шоумен', 'Планета обезьян: Революция']
```

Рис. 2. Рекомендация фильмов пользователю

Персонализированные рекомендации для пользователя с уникальным идентификатором 15647798 с сайта kinopoisk, изображены на рисунке 2.

Построенная рекомендационная система выводит фильмы в порядке убывания предсказанного значения. Чем больше значение, тем больше вероятность, что фильм придется пользователю по вкусу.

6. Метрики точности предсказания

Не существует универсальной метрики для оценки точности рекомендательной системы, поэтому в зависимости от целей и задач, метрики делятся на категории:

– оценка точности предсказываемого рейтинга;

– оценка релевантности рекомендаций;

– оценка качества ранжирования рекомендаций.

Поскольку фильмы оцениваются по непрерывной шкале (1-10), для оценки точности предсказываемого рейтинга рекомендательной системы подойдут регрессионные метрики, которые изображены в таблице 3:

r_i — фактическая оценка пользователя, $\hat{r}_i$ — предсказанная оценка, N — исходное количество фильмов.

Таблица 3

Регрессионные метрики

Название метрики	Формула	Описание
MAE (Mean Absolute Error)	$\frac{1}{N}\sum_{0}^{N-1}\lvert r_i - \hat{r}_i\rvert$	Среднее абсолютное отклонение
MSE (Mean Squared Error)	$\frac{1}{N}\sum_{0}^{N-1}\left(r_i - \hat{r}_i\right)^2$	Среднеквадратичная ошибка
RMSE (Root Mean Squared Error)	$\sqrt{\frac{1}{N}\sum_{0}^{N-1}\left(r_i - \hat{r}_i\right)^2}$	Корень из среднеквадратичного отклонения

Для вычисления MAE необходимо взять абсолютное значение разницы между предсказанной и фактической оценкой и вычислить их среднее значение [5].

MSE — ошибка вычисляется как среднеквадратичная разница между предсказанной и фактической оценкой.

RMSE — ошибка вычисляется как корень из среднеквадратичной разницы между предсказываемым значением и реальным значением.

Для созданной системы рекомендаций были построены функции вычисления метрик для ненулевых значений, результат работы программы приведен в таблице 4. Чем меньше значение ошибки, тем точнее прогноз. Чем больше сингулярных векторов используется при сингулярном разложении, тем точнее метрики.

Таблица 4

Влияние количества сингулярных чисел на точность

Название метрики	$k = 2000$	$k = 3681$	$k = 5000$	$k = 10000$
MAE	1.755	0.842	0.455	0.023
MSE	5.8	1.932	0.811	0.019
RMSE	2.41	1.39	0.9	0.14

Заключение

В результате были изучены методы снижения размерности. Рассмотрен пример, в котором реализовано сингулярное разложение, что позволило при 3681 сингулярных элементах, которые описывают 90 % дисперсии, сократить объем хранения данных в 3,4 раза. Поскольку в восстановленной матрице элементы зависят от сингулярных значений, была построена рекомендательная система, в которой элементы интерпретируются как степень важности предсказанных значений. В итоге была дана рекомендация фильмов для пользователя.

Список литературы

1. Вьюгин В. Математические основы теории машинного обучения и прогнозирования. – М.: МЦМНО, 2013. – 390 с.

2. Вержбицкий В.М. Вычислительная линейная алгебра: учебное пособие для вузов. – Изд. 3-е. – Москва; Берлин: Директ-Медиа, 2021. – 354 с.

3. Falk K. Practical recommender systems. – Shelter Island, NY: Manning, 2019. – 432 p.

4. Kaggle [Электронный ресурс]. – URL: https://www.kaggle.com/ (дата обращения: 15.10.22).

5. Chai T., Draxler R.R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature // Geoscientific Model Development. – 2014. – Vol. 7, No. 3. – Pp. – 1247–1250. – DOI: 10.5194/gmd-7-1247-2014.