КЛАСТЕРИЗАЦИЯ МНОГОМЕРНЫХ ОБЪЕКТОВ РАЗЛИЧНОЙ ПРИРОДЫ: ПОСТАНОВКА ИССЛЕДОВАТЕЛЬСКОЙ ЗАДАЧИ

В.М. Московкин, Казимиру Эринелту
г. Белгород, Россия, г. Луанда, Ангола

При кластеризации объектов различной природы целесообразно рассмотреть три вида объектов: скалярные, векторные и матричные. Google Scholar на запросы этих терминов на русском и английском языках (расширенный поиск, точная фраза) дает следующие количества откликов (табл. 1).

Таблица 1 Тестирование в Google Scholar терминов по различным видам объектов. 16.02.2017 г.

Термины	Количество откликов
Скалярный объект	920
Векторный объект	73
Матричный объект	4
Scalar object	888
Vector object	3460
Matrix object	4030

Как видим из таблицы 1, при усложнении объекта, встречаемость соответствующих русскоязычных терминов уменьшается, а англоязычных возрастает.

Можно дать следующие очевидные определения рассматриваемых объектов: скалярный объект – объект, состояние которого представлено скалярной величиной; векторный объект – объект, состояния которого представлены векторной величиной; матричный объект – объект, состояния которого представлены в матричном виде (в виде матрицы состояний). Очевидно, что многомерные объекты могут быть векторными или матричными.

Во многих областях знаний и на практике возникает задача распределения такого рода объектов на классы (группы). Чаще всего это делается для векторных объектов. Например, имеется n объектов, каждый из которых описывается m параметрами, тогда состояние i-го объекта описывается вектором $\overrightarrow{x_i} = (x_{1i}, x_{2i}, ..., x_j, ... x_{ni})$, где x_{ji} – значение j-го параметра для i-го объекта, $1 \le i \le n$, $1 \le i \le m$. Встаёт задача разбить эти объекты на группы таким образом, чтобы некая мера близости (расстояние, метрика) внутри этих групп объектов была минимальной, а мера близости между группами объектов максимальной. Эта типичная задача кластеризации, для решения которой используются различные методы и алгоритмы кластеризации.

В соответствии с тремя видами объектов напрашивается рассмотрение соответствующих терминов по их кластеризации. Тестирование этих терминов с помощью Google Scholar представлено в таблице 2.

Таблица 2
Тестирование в Google Scholar по различным типам кластеризации объектов 16.02.2017 г.

Термин	Количество откликов
Скалярная кластеризация	0
Векторная кластеризация	0
Матричная кластеризация	0
Scalar clustering	72
Vector clustering	4390
Matrix clustering	1920

Как видим, в отечественной литературе термины по типам кластеризации не используются.

В нашем понимании эти термины означают следующее:

1. Скалярная кластеризация (кластеризация скалярных объектов) – кластеризация объектов, состояния которых описываются скалярными величинами;
2. Векторная кластеризация (кластеризация векторных объектов) – кластеризация объектов, состояния которых описываются векторами одинаковой длины;

3. Матричная кластеризация (кластеризация матричных объектов) – кластеризация объектов, состояния которых описываются матрицами одинаковой размерности.

Наши эксперименты по тестированию этих терминов показывают, что первые работы по Scalar clustering появились в начале 80-х годов XX в., первые работы по Vector clustering с середины 60-х годов XX в., первые работы по Matrix clustering – с конца 60-х годов XX в. Отметим, что наше понимание векторной и матричной кластеризации отличается от большинства зарубежных работ. Очевидно, что все известные методы кластеризации относятся, в первую очередь, к кластеризации векторных объектов, то есть к векторной кластеризации в нашем понимании. В тоже время основной кластер публикаций, порожденный термином «Vector clustering» относится к векторному методу кластеризации под названием «support vector clustering» (SVC), в котором используются экспоненциальные Гауссовы ядра (Gaussian Rernel) [1]. Точно также под matrix clustering понимают не методологию кластеризации матричных объектов (матриц состояний объектов) одинаковой размерности, а различные методы кластеризации с использованием аппарата теории матриц. Например, в работах [3] под matrix clustering подразумевается новый data mining метод, который позволяет извлекать плотные суб-матрицы малой размерности из больших разряженных бинарных матриц.

В работе (June-Jei Kuo, Yu-Jung Zhang, 2012) со ссылкой на работу (S. Oyanagi, 2001) отмечается, что существует два класса алгоритмов кластеризации – Hierarchical clustering и Partional clustering, причем преимущества первого класса алгоритмов являются недостатками второго, и наоборот. В этой работе приводится следующая мысль: «Когда мы имеем дело с персонализацией данных встает вопрос о том, как учесть потребительские интересы, чтобы они стали важной исследовательской задачей». Согласно работе (S. Oyanagi, 2001) одним из подходов персонализации данных является matrix clustering, который подобен коллаборативной фильтрации (collaborative filtering). Он был приложен к Web access log посредством представления связи между пользователем и Web-страницами в бинарной матрице. Алгоритм такой матричной кластеризации называется «Ping-pong». Такого рода алгоритмы по извлечению плотных суб-матриц малой размерности из больших разряженных бинарных матриц используется в задачах оптимизации потребительских предпочтений (определение тесных групп клиентов, ориентированных на определенный круг товаров или услуг) [4, 5].

Существуют также и другие методы матричной кластеризации, например, Spectral Matrix clustering [2], отличные от нашего понимания процесса матричной кластеризации.

При этом суть кластеризации матричных объектов ничем не отличается от кластеризации векторных объектов: разбиение матричных объектов на группы таким образом, чтобы мера близости внутри выделенных групп матричных объектов была минимальной, а мера близости между выделенными группами матричных объектов максимальной. Поэтому для кластеризации объектов, представленных матрицами состояний одинаковой размерности, могут использоваться те же методы, которые используются и для кластеризации объектов, представленных векторами (метод Уорда, K-means метод и др.).

Исследовательская задача при кластеризации матричных объектов состоит в том, чтобы испытать различные меры близости (расстояния, метрики) и методы кластеризации, то есть необходимо понять в каких случаях различные метрики и методы кластеризации будут приводить к одним и тем же результатам, а в каких нет. Допустим, мы хотим кластеризовать n матричных объектов (n объектов, описываемых матрицами состояний одинаковой размерности) с помощью m метрик и k методов кластеризации, тогда нам нужно будет проделать $m\ k$ кластеризации (сценарных расчетов) и сравнить их результаты.

Отметим, что наш поиск в сети Интернет обнаружил небольшой ряд разрозненных задач в областях распознавания изображений, биоинформатике и экономике, в которых используются понятий расстояний между матрицами одинаковой размерности и методы кластеризации таких матриц. В тоже время отсутствует систематический подход для решения такого рода задач. Мы планируем его использовать для матричной кластеризации экономических объектов.

Литература

1. Asa Ben-Hur, David Horn, Hava T. Siegelmann and Vladimir Vapnik: Support Vector Clustering, Journal of Machine Learning Research 2, pp. 125-137, 2001.

2. Tian Zheng, Li Xiao Bin and Ju Yan Wei: Spectral clustering based on matrix perturbation theory, Sci China Ser F-Inf Sci, Vol. 50, No. 1, pp. 63-81, 2007.

3. Shigeru Oyanagi, Kazuto Kubota, and Akihiko Nakase: Mining WWW Access Sequence by Matrix Clustering, WEBKDD 2002, LNAI 2703, pp. 119–136, 2003.

4. S. Oyanagi, K. Kubota, A. Nakase: Application of Matrix Clustering to Web Log Analysis and Access Prediction, WEBKDD 2001.

5. S. Oyanagi, K. Kubota, A. Nakase: Matrix Clustering: A New Data Mining Method for CRM, Trans. IPSJ, Vol.42, No.8, pp. 2156–2166, 2001.

6. June-Jei Kuo, Yu-Jung Zhang: A Library Recommender System Using Interest Change over Time and Matrix Clustering, Taipei, Taiwan, pp. 259 - 268, 2012.