Система процедур визуального анализа многомерных данных

Бондарев А.Е., Галактионов В.А., Клышинский Э.С., Шапиро Л.З.
bond@keldysh.ru|vlgal@gin.keldysh.ru|klyshinsky@mail.ru|pls@gin.keldysh.ru
ИПМ им. М.В. Келдыша РАН

В работе рассматриваются задачи визуального анализа многомерных наборов данных. Для визуального анализа применяется подход построения упругих карт. В работе приведены результаты применения упругих карт для визуального анализа многомерных наборов данных различного происхождения. В частности, рассматривается задача анализа текстовой информации, представленной в виде многомерного массива частот совместного употребления глаголов и существительных. Описан ряд процедур обработки данных, позволяющих улучшить полученные результаты.

Ключевые слова: *многомерные данные, визуальный анализ, упругие карты.*

System of procedures for visual analysis of multidimensional data

Bondarev A.E., Galaktionov V.A., Klyshinsky E.S., Shapiro L.Z.
bond@keldysh.ru|vlgal@gin.keldysh.ru|klyshinsky@mail.ru|pls@gin.keldysh.ru
Keldysh Institute of Applied Mathematics RAS, Moscow, Russia

The paper considers the problems of visual analysis of multidimensional data. For visual analysis, the elastic maps approach is used. The paper presents the results of applying elastic maps for the visual analysis of multidimensional datasets having various origins. In particular, the problem of the analysis of textual information represented in the form of a multidimensional array of frequencies of joint use of verbs and nouns is considered. A number of data processing procedures are described that allow improving the results obtained.

Keywords: *multidimensional data, visual analysis, elastic maps.*

1. Введение

В анализе многомерных данных особое место занимают задачи классификации. При решении задач классификации весьма полезными оказываются подходы визуальной аналитики, являющиеся синтезом нескольких алгоритмов понижения размерности и визуального представления многомерных данных во вложенных в исходный объем многообразиях меньшей размерности.

К таким алгоритмам можно отнести отображение исходного многомерного объема в упругих картах [2,6,7] с разными свойствами упругости или эластичности. Эти методы позволяют тем или иным образом выделить из исходного многомерного объема данных содержащуюся в нем кластерную структуру. Авторами подхода [2,6,7] разработан программный комплекс ViDaExpert [3], позволяющий проводить построение и визуальное представление упругих карт. Основные функциональные особенности данного программного комплекса подробно описаны в [2].

Интерес к упругим картам появился у нас в процессе реализации проекта по разработке вычислительной технологии для построения, обработки, анализа и визуального представления многомерных параметрических решений задач газовой динамики. Вычислительная технология реализована как единая технологическая цепочка алгоритмов производства, обработки, визуализации и анализа многомерных данных. Такая технологическая цепочка может рассматриваться как прототип обобщенного вычислительного эксперимента для нестационарных задач вычислительной газовой динамики. В итоге подобный обобщенный вычислительный эксперимент позволит получать решение не одной отдельно взятой задачи, а решение для целого класса задач, задаваемого диапазонами изменения определяющих параметров. Также следует отметить универсальность подобного обобщенного вычислительного эксперимента. Он может быть применен к широкому кругу задач математического моделирования нестационарных процессов. Практическая реализация подобного обобщенного эксперимента может обеспечивать организацию крупномасштабных промышленных расчетов. Описание элементов реализованной вычислительной технологии приведено в работах [4,5].

На практике упругие карты оказались полезным и достаточно универсальным инструментом, что позволило применять их к многомерным объемам данных разного типа. Данный подход был применен к задачам анализа текстовой информации, где в качестве числовых характеристик выступали частоты употребления слов [1].

2. Упругие карты

Идеология и алгоритмы реализации построения упругих карт подробно представлены в работах [2,6]. Подобная карта представляет собой систему упругих пружин, вложенную в многомерное пространство данных. Этот подход основывается на аналогии с задачами механики: главное многообразие, проходящее через «середину» данных, может быть представлено как упругая мембрана или пластинка. Метод упругих карт формулируется как оптимизационная задача, предполагающая оптимизацию заданного функционала от взаимного расположения карты и данных.

Согласно [2], основой для построения упругой карты является двумерная прямоугольная сетка G, вложенная в многомерное пространство, которая аппроксимирует данные и обладает регулируемыми свойствами упругости по отношению к растяжению и изгибу.

Варьирование параметров упругости заключается в построении упругих карт с последовательным уменьшением коэффициентов упругости, в силу чего карта становится более мягкой и гибкой, наиболее оптимальным образом подстраиваясь к точкам исходного многомерного объема данных. После построения упругую карту можно развернуть в плоскость для наблюдения кластерной структуры в изучаемом объеме данных. Применение упругих карт позволяет более точно и четко определять кластерную структуру изучаемых многомерных объемов данных.

Следует отметить, что при построении упругих карт в многомерном облаке данных, состоящем из сгущений и

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

отдельных отдаленных точек, возникает проблема масштабируемости. Упругая карта будет пытаться подстроиться под рассматриваемый объем в целом – как к отдаленным точкам, так и к областям сгущения, что, естественно, не может получиться одинаково хорошо. Для того чтобы решить эту проблему и обеспечить четкое представление о данных в области сгущений в работе [1], был предложен подход, названный quasi-Zoom, заключающийся в вырезании области сгущения из рассматриваемого облака многомерных данных и построения для вырезанной области упругой карты заново.

3. Процедуры обработки многомерных данных при построении упругих карт

Рассмотрим пример построения упругих карт для объема многомерных данных, представляющих собой описание характеристик полезных ископаемых, а именно, трех сортов угля из месторождений Польши [8].

Рассматриваются многомерные данные, представляющие собой точки в многомерном пространстве признаков (характеристик образцов угля). Пространство признаков состоит из следующих характеристик образцов угля – плотность, масса, удельная теплота сгорания, зольность, содержание серы, содержание летучих компонент, содержание влаги.

Таким образом, мы имеем набор точек в 7-мерном пространстве, соответствующих различным образцам угля. В наборе данных отображены три сорта угля. Рассматривается визуальный анализ с помощью применения упругих карт и главных компонент с целью изучения кластеризации многомерного облака данных и разделения сортов угля. Здесь и далее построение и визуальное представление упругих карт реализовано с помощью программного комплекса ViDaExpert [3], подробно описанного в [2].

Для исходного объема строится «мягкая» упругая карта, отображаемая в пространстве, образованном первыми тремя главными компонентами. Красные, зеленые и синие точки соответствуют трем типам угля (см. рис. 1). Далее представляем развертку построенной карты (см. рис. 2) на плоскость, образованную двумя первыми главными компонентами.

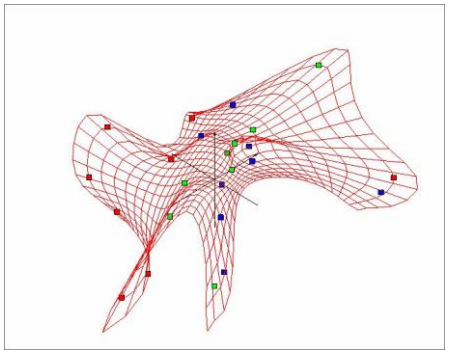

Рис. 1. Построение «мягкой» упругой карты, представляющей три сорта угля.

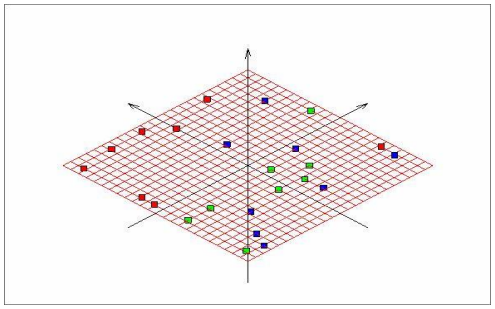

Рис. 2. Развертка «мягкой» упругой карты, представленной на предыдущем рисунке.

По развертке видно, что красные точки образуют отделившийся кластер, за исключением точки в правой части рисунка. Зеленые и синие точки перемешаны.

Для дальнейшего улучшения разделения применим фильтрацию исходного объема данных. Следует отметить, что для некоторых точек в исходном массиве представлены неполные данные, то есть для некоторых образцов информация по ряду характеристик отсутствует или находится в широком диапазоне вариации, а не представлена точно. В частности, данные по размерам образцов представлены неопределенной величиной меньше некоторого или больше некоторого предела (огромные куски или пыль). Попробуем провести фильтрацию данных, то есть убрать все точки, данные по которым представлены нечетким или неполным образом.

Удаление подобных точек из исходного объема приводит к следующим результатам (см. рис. 3).

На рисунке 3 представлена развертка «мягкой» упругой карты для измененного объема данных.

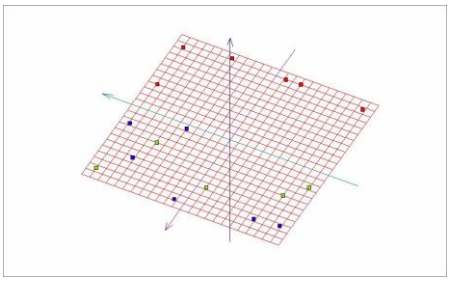

Рис. 3. Развертка «мягкой» упругой карты после фильтрации данных.

Сорт угля, представленный точками красного цвета, полностью отделился после процедуры удаления из исходного объема данных точек с нечетко определенными координатами в 7-мерном пространстве.

Сорта, представленные синими и зелеными точками, остались смешанными. Попробуем еще раз провести ту же процедуру удаления точек из данных. Однако на этот раз исключим из рассматриваемого объема красные точки целиком. Назовем эту процедуру флотацией (от английского термина flotation) аналогично термину, применяющемуся при очистке горных пород, когда более легкие фракции всплывают на поверхность и удаляются.

Теперь для четкого разделения двух оставшихся сортов достаточно отобразить точки нового объема данных в пространстве трех первых главных компонент (см. рис. 4).

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18

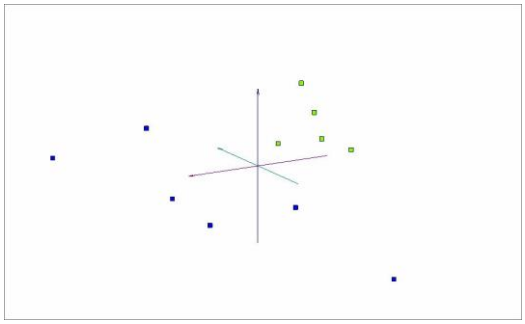

Рис. 4. Оставшиеся точки в пространстве трех первых главных компонент после флотации (удаления отделившегося полностью класса).

Итак, комбинируя построение упругих карт, отображение на вложенные в исходное пространство главные 2D и 3D многообразия и операции удаления нечетких точек и отделившихся классов (фильтрация и флотация данных), можно полностью провести разделение заданных в исходном объеме образцов на три класса, соответствующих трем сортам угля.

Эти же приемы обработки исходных данных послужили основой для разработки приема quasi-Zoom при анализе текстовой информации в работе [1]. В этой работе основное внимание уделялось изучению возможности применения методов упругих карт и t-SNE (метод стохастического распределения соседей с использованием t-распределения Стьюдента) для анализа тематической близости слов русского языка.

Основой предлагаемого метода является анализ непосредственного окружения слов. Основная гипотеза состоит в том, что близкие по смыслу слова должны встречаться в примерно одинаковом контексте. В связи с этим в пространстве признаков они будут находиться на относительно близком расстоянии друг от друга, тогда как отличающиеся слова будут находиться на более удаленном друг от друга расстоянии.

Рассмотрим результаты построения упругих карт для тестового объема [1]. В качестве тестовых данных использовались текстовые корпуса из новостных источников (ленты новостей за определенный период). Процедуры извлечения синтаксически связанных пар слов подробно описаны в [1]. Для первичных тестов было отобрано около 100 глаголов со 155 наиболее связанными с ними существительными. Полученные таким образом данные далее рассматривались как многомерный объем данных, представляющий собой 100 точек в 155-мерном пространстве. Числовые значения получающейся в результате матрицы определяются как частоты совместного употребления. Рассмотрим применение упругих карт к изображению данного объема. На рисунке 5 представлена развертка «мягкой» упругой карты для данного объема.

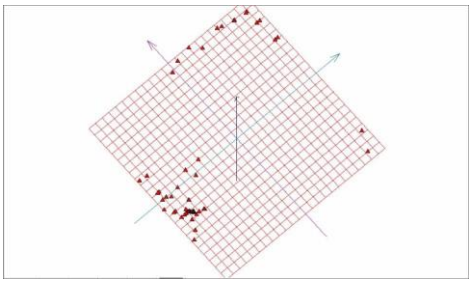

Рис. 5. Построение развертки «мягкой» упругой карты на плоскость первых двух главных компонент [1].

На рисунке видно, что изучаемый объем данных содержит область высокой плотности данных и точки, достаточно далеко отстоящие от этой области.

Именно в таких случаях возникает проблема масштабируемости, описанная ранее. Для решения этой проблемы был разработан подход quasi-Zoom, представленный в работе [1]. При исследовании объема частот совместного употребления глаголов и существительных в работе [1] практическая задача ставилась следующим образом. Нужно было максимально разделить «слипшиеся» точки. Для этой цели был разработан подход quasi-Zoom, который позволил решить эту задачу. Суть этого технологического приема заключается в том, что для более тонкой подстройки необходимо выделять большие кластеры в исследуемом объеме многомерных данных и проводить построение упругих карт для выделенных кластеров отдельно, организуя тем самым эффект, подобный функции «zoom» в современной фототехнике. Это позволит избежать проблем с масштабируемостью, когда упругая карта должна описывать как области сгущения, так и сильно удаленные отдельные точки. Этот прием сродни проиллюстрированным ранее фильтрации и флотации, однако он отличается от них тем, что имеет четкую последовательность применения.

Для улучшения разделения в области сгущения применим подход quasi-Zoom. Из исходного многомерного объема данных вырежем отделившиеся точки верхней части развертки. К получившемуся в результате этой процедуры новому объему многомерных данных заново применим построение упругой карты. При этом упругую карту будем строить сразу максимально мягкой.

После первого применения подхода quasi-Zoom разделение точек удалось значительно улучшить. Однако под точкой, соответствующей наивысшей плотности данных скрываются еще 37 «слипшихся» слов – глаголов, приведенных на рисунке 6.

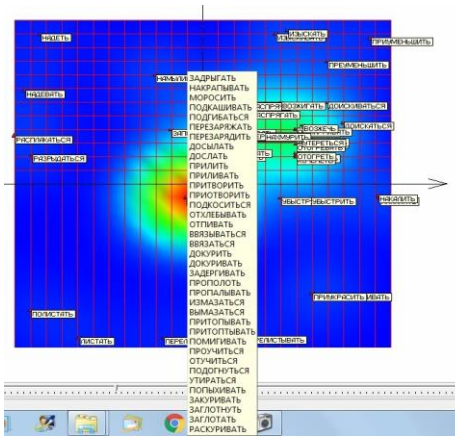

Рис. 6. Список «слипшихся» глаголов в точке наивысшей плотности данных после первого применения подхода quasi-Zoom с раскраской по плотности.

Для того, чтобы обеспечить разделение оставшихся 37 глаголов в точке наивысшей плотности данных, применим прием quasi-Zoom еще раз. Снова вырезаются отделившиеся точки таким образом, чтобы в результирующем многомерном объеме данных остались только 37 «слипшихся точек».

Повторно проводим для получившегося нового многомерного объема данных из 37 точек построение максимально мягкой упругой карты из сетки с таким же

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

количеством узлов, как и на предыдущем этапе. Большую часть точек удалось разделить, однако в зоне наивысшей плотности данных осталось еще 17 «слипшихся» точек.

Проводим процедуру quasi-Zoom в третий раз для оставшихся 17 точек. Результаты представлены на рисунке 7. В результате третьего применения процедуры удалось отделить еще 4 точки. Непосредственный анализ частот встречаемости для оставшихся 13 точек показал, что все они имеют одинаковые координаты по всем измерениям равные нулю. Следовательно, их разделение невозможно в принципе.

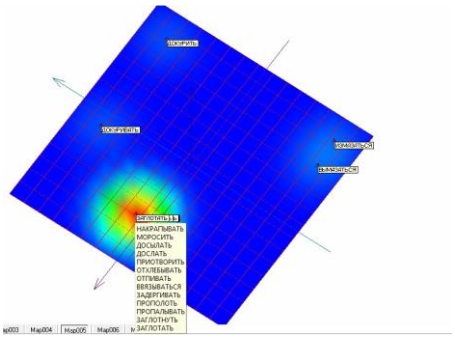

Рис. 7. Развертка упругой карты с раскраской по плотности после третьего применения процедуры quasi-Zoom [1].

Таким образом, процесс разделения точек в тестовом многомерном объеме данных с помощью построения упругих карт и применения процедуры quasi-Zoom завершен полностью и успешно. Удалось разделить все точки многомерного объема данных, имеющие различные и отличающиеся от нуля координаты. Апробация приема quasi-Zoom, проведенная в работе [1] показала эффективность разработанного подхода. При этом результаты показали необходимость проведения фильтрации данных на предварительном этапе с целью удаления точек с полностью нулевыми или полностью одинаковыми координатами. Для этой цели был построен автоматический фильтр в виде программного модуля, позволяющий проводить фильтрацию по указанному признаку. Для полного разделения точек рассматриваемого объема потребовалось только два раза последовательно применить quasi-Zoom. Повторное применение quasi-Zoom уже не оставляет слипшихся точек. Таким образом, можно сделать вывод, что предварительная очистка данных путем фильтрации существенно ускоряет обработку многомерного массива и улучшает качество обработки.

Следующим интересующим нас вопросом являлся вопрос влияния на результат транспонирования рассматриваемого многомерного массива данных. Что будет, если поменять существительные и глаголы местами? Теперь рассмотрим 155 точек, характеризующих существительные, а в качестве измерений у нас будут служить 100 глаголов (87 после фильтрации). Соответственно многомерный массив будет также состоять из частот совместного употребления существительных и глаголов, и мы сможем посмотреть и оценить, насколько схожие в понятийном смысле слова будут близки на развертке упругой карты. На рисунках 8 и 9 представлены верхняя правая и нижняя правая части развертки упругой карты. На рисунке 8 можно видеть достаточно четко выделенный понятийный кластер – ЧЕЛОВЕК, ЖЕНЩИНА, ДЕВОЧКА, ДЕВУШКА, МАТЬ, РЕБЕНОК. Чуть выше менее четко просматривается смысловой кластер – РУКА, ПАЛЕЦ, ЛИЦО, ТЕЛО.

На рисунке 9 также просматривается кластер понятий – СМЫСЛ, ДОКУМЕНТ, БУМАГА, ГАЗЕТА, КНИГА, РОМАН, ОТВЕТ.

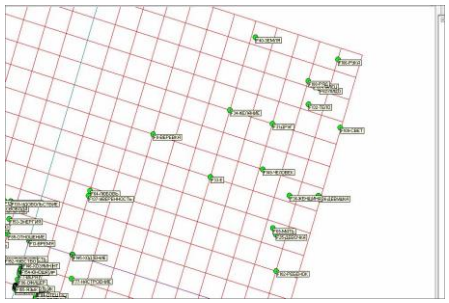

Рис. 8. Верхняя правая часть развертки упругой карты.

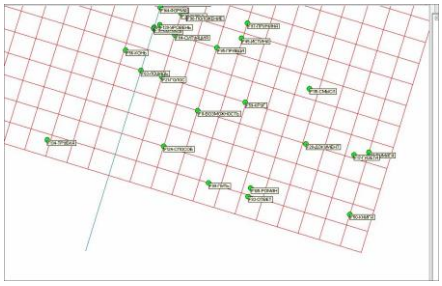

Рис. 9. Нижняя правая часть развертки упругой карты.

Это открывает широкие возможности по анализу и трактовке смысловых группировок для специалистов в этой области. Для более тщательного анализа и трактовки смысловой близости следует развивать данные подходы с точки зрения анализа расстояний между точками многомерного массива для различного задания метрики. Также интересной и важной задачей в этом плане должно стать выяснение влияния выбора метрики.

Для улучшения разделения необходимо в дальнейшем усовершенствовать алгоритм построения упругих карт за счет возможности сгущения исходной сетки карты в областях повышенной плотности данных. Подобная возможность широко используется в задачах математического моделирования сплошных сред.

4. Заключение

Для анализа кластерных структур в многомерных объемах данных использованы технологии построения упругих карт, представляющие собой методы отображения точек исходного многомерного пространства на вложенные многообразия меньшей размерности. Рассмотрен ряд приемов обработки данных, позволяющих улучшить результаты – предварительная фильтрация данных, удаление отделившихся кластеров (флотация). Для решения проблемы масштабируемости, когда упругая карта подстраивается как под область сгущения точек данных, так и к отдельно расположенным точкам облака данных, применяется подход quasi-Zoom. Суть подхода заключается в том, что для более тонкой подстройки в многомерном облаке данных выделяются большие кластеры, после чего проводится отдельное построение упругих карт для выделенных кластеров. Приведены примеры построения упругих карт для многомерных данных различного происхождения.

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18

5. Благодарности

Данная работа выполнена при поддержке Российского фонда фундаментальных исследований (проекты 16-01-00553а и 17-01-00444а).

6. Литература

[1] Бондарев А.Е., Бондаренко А.В., Галактионов В.А., Клышинский Э.К. Визуальный анализ кластерных структур в многомерных объемах текстовой информации / Научная визуализация. 2016. Т.8. № 3. с. 1-24.

[2] Зиновьев А.Ю. Визуализация многомерных данных - Красноярск: Изд-во КГТУ, 2000 - 168 с.

[3] Программный пакет ViDaExpert http://bioinfo.curie.fr/projects/vidaexpert/ (дата обращения 01.02.2018).

[4] Bondarev A.E., Galaktionov V.A. Analysis of Space-Time Structures Appearance for Non-Stationary CFD Problems / Proceedings of 15-th International Conference On Computational Science ICCS 2015 Rejkjavik, Iceland, June 01-03 2015, Procedia Computer Science. Vol. 51. P. 1801–1810.

[5] Bondarev A.E., Galaktionov V.A. Multidimensional data analysis and visualization for time-dependent CFD problems / Programming and Computer Software. 2015. Vol. 41. №. 5. P. 247–252. DOI: 10.1134/S0361768815050023

[6] Gorban A., Kegl B., Wunsch D., Zinovyev A. (Eds.), Principal Manifolds for Data Visualisation and Dimension Reduction, LNCSE 58, Springer, Berlin – Heidelberg – New York, 2007.

[7] Gorban A.N., Zinovyev A. Principal manifolds and graphs in practice: from molecular biology to dynamical systems / International Journal of Neural Systems. 2010. Vol. 20. №. 3. P. 219–232. DOI: 10.1142/S0129065710002383

[8] Niedoba T. Multi-parameter data visualization by means of principal component analysis (PCA) in qualitative evaluation of various coal types / Physicochemical Problems of Mineral Processing. 2014. Vol. 50. № 2. P. 575-589.

Об авторах

Бондарев Александр Евгеньевич, к.ф.-м.н., старший научный сотрудник, ИПМ им. М.В. Келдыша РАН.
Его e-mail bond@keldysh.ru.

Галактионов Владимир Александрович, д.ф.-м.н., профессор, зав. отделом, ИПМ им. М.В. Келдыша РАН. Его e-mail vlgal@gin.keldysh.ru.

Клышинский Эдуард Станиславович, к.т.н, доцент, НИУ ВШЭ МИЭМ. Его e-mail klyshinsky@mail.ru.

Шапиро Лев Залманович, к.т.н, старший научный сотрудник, ИПМ им. М.В. Келдыша РАН. Его e-mail pls@gin.keldysh.ru.

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18