Разработка математических моделей и алгоритма для решения задачи анализа многомерных динамических данных методом визуализации*

***Д.Д. Попов*[1], *И.Е. Мильман*[1], *В.В. Пилюгин*[1], *А.А. Пасько*[2]**

dpopovmephi@gmail.com|igalush@gmail.com|VVPilyugin@mephi.ru|apasko@bournemouth.ac.uk

Москва, Россия, [1]Национальный исследовательский ядерный университет «МИФИ»;

Борнмут, Великобритания, [2]Британский национальный центр компьютерной анимации при университете Борнмута

Статья посвящена анализу изменения объектов, характеризующихся набором из n численных параметров. Исследуются многомерные динамические данные об этих объектах в геометрической интерпретации. Инструментом проведения анализа является метод визуализации.

Рассмотрены различные подходы к визуализации многомерных данных. Приведено формальное описание метода решения задачи. Разработаны математические модели каждого этапа визуализации исходных данных.

Предложен алгоритм решения задачи. Описана разработанная интерактивная прикладная программа визуализации, реализующая алгоритм.

Отмечается эффективность использования метода визуализации. Разработанный на его основе алгоритм позволяет строить суждения об образовании и разрушении кластеров и сгустков из объектов, формально представленных в виде n-мерных наборов действит ельных чисел. Также алгоритм позволяет находить объекты, стремящиеся оказаться в кластере или сгустке или наоборот покинуть их. Приведены примеры использования программы. Показано, как с её помощью можно проводить макроанализ кредитных организаций, а такж е искать инварианты в изменении исходных данных.

Ключевые слова: *многомерный анализ, анализ динамических данных, визуальный анализ, многомерный визуальный анализ.*

1. Введение

Визуальная аналитика — это искусство аналитически рассуждать, поддерживаемое интерактивным визуальным интерфейсом [1]. Этот термин ввёл Джеймс Томас (*James Thomas*) в [3], чем обратил внимание научной общественности на роль визуального представления данных в решении исследовательских задач.

В данной статье предлагается метод решения класса задач анализа числовых данных, представляемых в табличном виде, данных о некотором изменении состояния рассматриваемой совокупности объектов во времени.

Методологической основой работы является сформулированный в [10] метод научной визуализации.

2. Постановка задачи

В геометрической интерпретации исследуется процесс изменения взаимного расположения многомерных точек из заданного множества. Под точкой понимается n-мерный набор действительных чисел, которые могут быть значениями параметров конкретного объекта. Точки множества являются элементами евклидова пространства E^n. В этом пространстве могут быть выделены подмножества точек, составляющие сгустки и кластеры . Определения этих подмножеств даны в [9].

Координаты точек могут изменяться с течением времени.

Решение задачи анализа изменения взаимного расположения точек заданного множества методом визуализации включает следующие этапы:

- Создание математических моделей исходных данных, операций над исходными данными, их отображения.
- Разработка алгоритма решения задачи и способов представления объектов и явлений.
- Написание прикладной программы.

При изложении последующего материала будем пользоваться понятием *геометрического процесса*. *Геометрический процесс* — это множество точек пространства, координаты которых зависят от времени.

3. Дискретный геометрический процесс

Исходными объектом анализа является множество n-мерных точек $P = \{p_1, p_2, \ldots, p_m\}$, значения координат которых заданы для нескольких моментов времени $T = \{t_1, t_2, \ldots, t_k\}, t_1 < t_2 < \cdots < t_k$.

$$p_i = \left(p_i^1, p_i^2, \ldots, p_i^n\right)$$

Для каждой пары точек определено $\rho\left(p_i, p_j\right) \geqslant 0$ — расстояние между точками p_i и p_j $\left(i, j = \overline{1, m}\right)$.

Работа опубликована при финансовой поддержке РФФИ, грант №16-07-20482

[1](англ.) Visual Analytics is the science of analytical reasoning supported by interactive visual interfaces

Таким образом изначально задан дискретный геометрический процесс $P = P(t_j)$, представляющий собой множество, $p_i = p_i(t_j)$ — n-мерные точки, каждая из которых, в свою очередь задаётся совокупностью $p_i^l = p_i^l(t_j)$ $\left(l = \overline{1,n}, i = \overline{1,m}, j = \overline{1,k}\right)$ — координаты точек.

4. Преобразования процессов

Введём ряд преобразований исходного дискретного геометрического процесса.

$P(t_j)$ представляет собой упорядоченный по времени набор описаний известных состояний множества исследуемых объектов. Сами же объекты существуют и изменяются непрерывно. Исходя из этого факта, необходимо иметь инструмент преобразования дискретного процесса в непрерывный. Таким инструментом служит интерполяция.

4.1 Интерполяция

Задача интерполяции состоит в поиске такой функции F из заданного класса функций, что $F(t_j) = P(t_j)$, где $t_j \in T$.

Так как

$$P = P(t_j) = \{p_1(t_j), p_2(t_j), \ldots, p_m(t_j)\},$$

$$p_i(t_j) = \left(p_i^1(t_j), p_i^2(t_j), \ldots, p_i^n(t_j)\right),$$

то поиск F заключается в поиске таких f_i^l, принадлежащих заданному классу, что $f_i^l(t_j) = p_i^l(t_j)$, $\left(i = \overline{1,m}, l = \overline{1,n}, j = \overline{1,k}\right)$. Пусть найдены непрерывные f_i^l, тогда, можно сказать, что найдена F. Непрерывный геометрический процесс, полученный в результате интерполяции исходного дискретного геометрического процесса $P(t_j)$, будет обозначать $P(t)$.

4.1.1 Кусочно-линейная интерполяция

f_i^l является алгебраическим двучленом $f_i^l = a_h t + b_h$ на каждом интервале $[t_h, t_{h+1}], 1 \leqslant h \leqslant k-1$. В таком случае при любом $t \in [t_h, t_{h+1}], 1 \leqslant h \leqslant k-1$, f_i^l будет рассчитываться по формуле:

$$f_i^l(t) = p_i^l(t_h) + \frac{p_i^l(t_{h+1}) - p_i^l(t_h)}{t_{h+1} - t_h}(t - t_h)$$

4.2 Дискретизация геометрических процессов

Пусть $P(t)$ — непрерывный геометрический процесс. Тогда $P(t_0)$ — его временное сечение, t_0 — принадлежит области определения $P(t)$.

Из $P(t)$ можно построить дискретный геометрический процесс, заданные в k' моментах времени. Для этого выбираем $\tau_1, \tau_2, \ldots, \tau_{k'}$ и определим $P'(\tau_h)$ как сосвокупность временных сечений $P(t)$: $P'(\tau_h) := \{P(\tau_h) | \tau_h \in \{\tau_1, \tau_2, \ldots, \tau_{k'}\}\}$.

5. Подходы к визуальному анализу многомерных динамических данных

5.1 Особенности зрительного восприятия человека в контексте решения задач анализа данных визуальным методом

Задача средства визуализации – возникновение инсайта у пользователя (англ. insight – проницательность, понимание, озарение, интуиция). Это может выступать критерием оценки качества разработанного продукта. Об этом пишут в [3, 2, 5, 6] , [8].

Приведём некоторые замечания из [12], которые могут быть использованы при разработке программ визуализации.

Наиболее информативным является кодирование данных формой. При этом о сновное значение в восприятии формы человеком имеет отношение «фигура-фон». Эмпирически выявлены следующие принципы выделения фигур и фона на рассматриваемом изображении:

- В качестве фигуры, прежде всего, выделяются замкнутые конфигурации.
- Симметричные конфигурации легче воспринимаются как фигуры, чем конфигурации ассиметричные.
- В том случае, когда поле изображения заполнено однородными элементами, фигуру образуют те из них, которые расположены ближе друг к другу.
- Если поле изображения заполнено разнородными элементами, то фигура образуется, прежде всего, теми из них, которые имеют сходство по форме или цвету.
- Если те или иные элементы перемещаются по полю изображения в одном направлении и с одинаковой скоростью, то именно они выделяются как фигура.
- Если расположить часть элементов в определённом порядке, то можно создать у наблюдателя установку, которая повлияет на восприятие остальных элементов.

Опираясь на вышеприведённые данные, можно оценить возможности уже имеющихся программных продуктов, осуществляющих визуализацию многомерных данных. Кроме того, рассмотренные наблюдения целесообразно использовать и при их разработке.

5.2 Средства визуализации динамических многомерных данных

Примером визуализации многомерных динамических данных служит средство контроля состояния пациента, описанное в [11]. В указанной статье производится анализ одного объекта, характеризующегося множеством числовых параметров (мощ-

ность этого множества > 3), значения которых изменяются с течением времени.

В [7] описывается опыт раз работки подсистемы визуализации, основная цель которой — индикация возникновения критического состояния наземных станций командно-измерительных систем (КИС). Индикация осуществляется путём изменения цвета или формы элементов пользовательского интерфейса системы контроля КИС.

В работе [1] описываются подходы к решению задачи анализа динамической системы с использованием визуализации. Динамическая система задаётся набором дифференциальных уравнений. Системе ставится в соответствие точка n–мерного евклидова пространства E_n, где n — число дифференциальных уравнений.

В работе [4] рассматриваются подходы к визуализации и анализу данных, формируемых автоматически программами - видеоиграми. Указанные данные также имеют временную составляющую. Однако, специфика их происхождения оказывает большое влияние на методы их анализа, рассматриваемые в статье.

6. Формализация метода решения задачи

Метод визуализации подразумевает последовательное решение двух задач, представленных на рисунке 1.

Задание параметров визуализации и получение изображений или временных последовательностей кадров происходит до тех пор, пока аналитик не рассмотрит достаточное количество для формирования некоторого суждения об изменении взаимного расположения заданного множеств а точек с течением времени.

Рассмотрим подробнее процесс визуализации исходных данных.

6.1 Задание динамических исходных данных

Задаётся дискретный геометрический процесс $P = P(t_j)$, для которого будет проводиться визуализация.

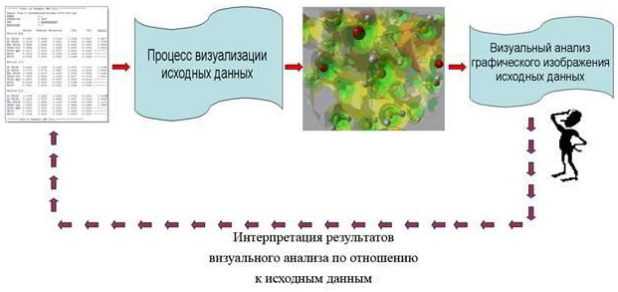

Рис. 1: Метод визуализации

6.2 Фильтрация

Для решения поставленной задачи исходный дискретный процесс $P = P(t_j)$ интерполируется линейно, как описано в 4.1. В результате фильтрации получается непрерывный процесс $P(t)$.

6.3 Мэппинг

На этом этапе исходным данным ставится в соответствие динамическая пространственная сцена $S(t) =< G(t), O(t) >$, где $G(t)$ — описание геометрии сцены, а $O(t)$ — оптических параметров сцены.

В каждый заданный момент t, $S(t)$ соответствует $P(t), t \in [t_1, t_k]$.

$$G(t)=\{Sph_1(t),Sph_2(t),...,Sph_m(t),Cyl_1(t),Cyl_2(t),...,Cyl_e(t)\}$$

где $Sph_i(t)$ — сфера соответствующая точке $p_i(t)$; $Cyl_j(t)$ — цилиндр, соединяющий 2 сферы, $e = C_m^2 = \frac{m!}{(m-2)!2!} = \frac{m(m-1)}{2}$.

$$O(t)=\{SphC_1(t),...,SphC_m(t),CylC_1(t),...,CylC_e(t)\}$$

где $SphC_1(t) \equiv \cdots \equiv SphC_m(t) \equiv SphC \equiv const$ — цвета сфер, $CylC_1(t) \ldots, CylC_e(t)$ — цвета цилиндров.

$$CylC_l(t) = \{Red(t), Green(t), Blue(t), Opacity(t)\},$$

$$Red(t) = 255\left(1 - \frac{\rho(p_{l_1}(t), p_{l_2}(t))}{d}\right)$$

$$Green(t) = 150\frac{\rho(p_{l_1}(t), p_{l_2}(t))}{d}$$

$$Blue(t) = 255\frac{\rho(p_{l_1}(t), p_{l_2}(t))}{d}$$

$$Opacity(t) = \begin{cases} 0, \text{при } \rho\left(p_{l_1}(t), p_{l_2}(t)\right) > d \\ 100, \text{иначе} \end{cases},$$

где значение $Opacity = 0$ означает, что цилиндр абсолютно прозрачен, $Opacity = 100$ — абсолютно непрозрачен, $Red, Green, Blue \in [0, 255]$.

6.4 Рендеринг

Результатом рендеринга является проекционное изображение I сцены S.

Каждому моменту времени t соответствует своя сцена, а значит и своё проекционное изображение:

$$I(t) = I(S(t), A(t)),$$

$A(t)$ – атрибуты визуализации. Например, камера,освещение, физические характеристики среды, в которой находится сцена, размер получаемого изображения.

7. Алгоритм решения задачи

Разработанный алгоритм решения поставленной задачи изображён на следующем рисунке.

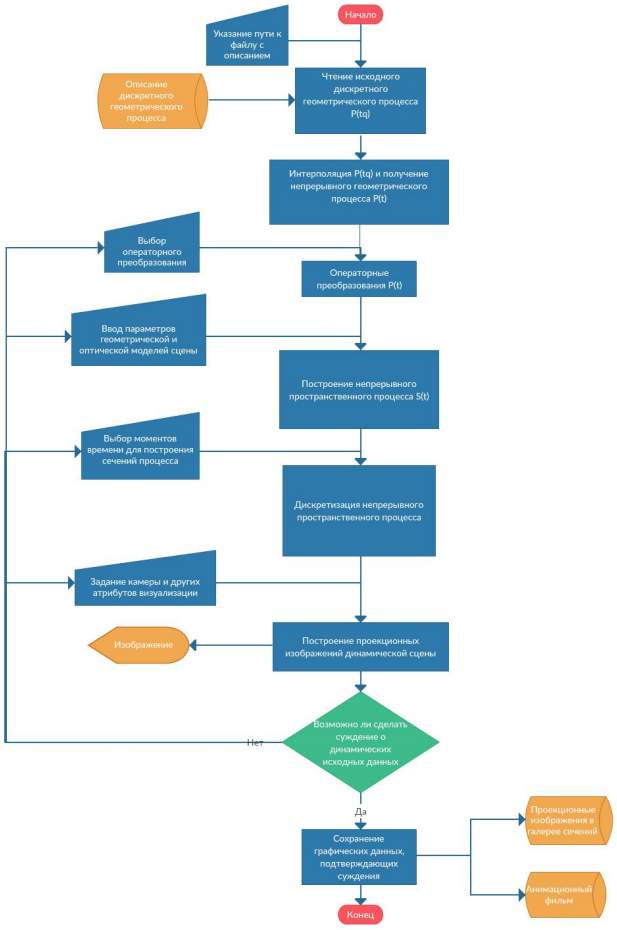

Рис. 2: Алгоритм решения задачи

8. Пример использования разработанной программы

С помощью созданной программы был произведён макроанализ 81 кредитной организации. Данные представляли собой ежемесячные отчёты этих организаций по 9 показателям за 13 месяцев 2013-2014 гг. Данные за первый отчётный месяц соответствуют $t = 1$, за последний — $t = 13$.

Под макроанализом понимается изучение принадлежности кредитных организаций в их геометрической интерпретации к подмножествам из 2. Отнесение точек к кластерам или сгусткам позволяет судить о схожести, подобии исследуемых объектов по интегральному показателю, заданному как расстояние между точками пространства E^n. На рисунке приведён ряд изображений, позволяющих сделать вывод, что за промежуток времени $t \in [7; 8]$ удалённая точка, соответствующая выделенной красным сфере, присоединилась к сгустку. Было замечено, что для одной из построенных моделей сцен наблюдается следующая закономерность. Большинство сфер, за исключением отмеченных зелёным на рисунке, лежат в одной плоскости. Это позволило выделить зависимость между тремя финансовыми показателями организаций:

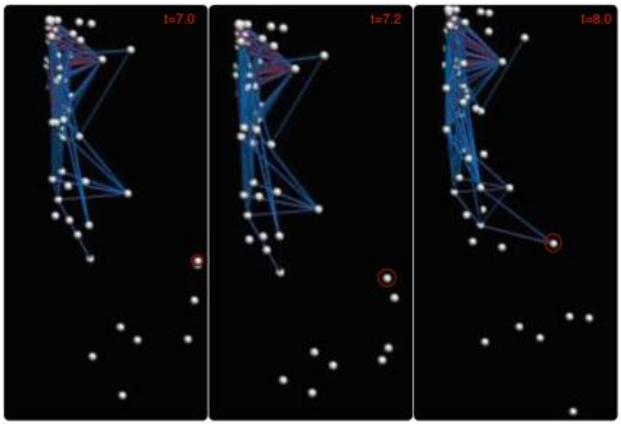

Рис. 3: Кадры полученного анимационного фильма

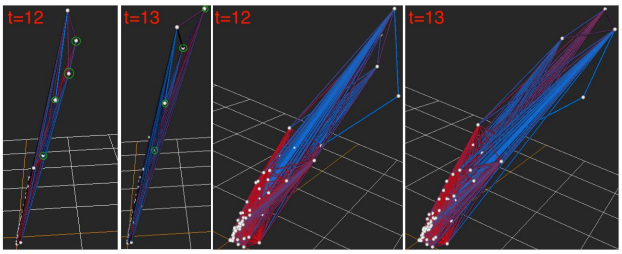

Рис. 4: Проекционные изображения пространственной сцены

X — прощёная задолженность в кредитном портфеле,
Y — выпущенные облигации,
Z — активы нетто.
Она описывается уравнением плоскости:

$$2,231X - 72,672Y + 20,3624Z - 2,58513 = 0.$$

Выделенные зелёным сферы и соответствующие им кредитные организации этой зависимости не подчиняются. Зависимость имеет место при $t \in [1; 9] \cup [12; 13]$.

9. Заключение

В этой работе:

- Были описаны, создан ные математические модели исходных данных, операций над ними.
- Приведён разработанный алгоритм решения задачи.
- С помощью программы, в основу которой лёг разработанный алгоритм, проведён анализ кредитных организаций.

В дальнейшем планируется расширение функциональных возможностей программы: выбор способа интерполяции процесса, дополнительные построения в пространственной сцене, а именно трёхмерные поверхности, аппроксимирующие исходные данные.

Литература

[1] Fischel G. Case Study: Visualizing Various Properties of Dynamical Systems / G. Fischel [et al.]. // In

Proceedings of the Sixth International Workshop on Digital Image Processing and Computer Graphics (SPIE DIP-97). – 1998. – vol. 3346. – P. 146-154.

[2] Haghverdi L., Buettner F., Theis F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data // Bioinformatics. – 2015. – №31(18). – P. 2989–2998.

[3] Thomas J., Cook K. Illuminating the Path: Research and Development Agenda for Visual Analytics. – IEEE-Press, 2005. – 185 p..

[4] Wallner G., Kriglstein S. PLATO: A visual analytics system for gameplay data // Computers & Graphics. – 2014. – №38. – P. 341-356.

[5] Авербух В. Л., Манаков Д. В. Анализ и визуализация "больших данных" // Труды междунар. науч. конф. "Параллельные Вычислительные Технологии" (ПаВТ'2015). Екатеринбург, 31 марта - 2 апреля 2015. – Челябинск, Издательский центр ЮУрГУ, 2015. – С. 332-340.

[6] Барсегян, А. А. Анализ данных и п роцессов: учеб. пособие / А. А. Барсегян [и др.]. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.

[7] Емельянова Ю. Г. Средства когнитивной графики для отображения и анализа т екущего состояния наземных станций командно-измерительных систем // Сб. трудов Инст. прогр. систем РАН и «Университета г. Переславля» им. А. К. Айламазяна. – Переславль-Залесский: Изд -во «Университет города Переславля», 2008. – том 1. – С. 111-121.

[8] Кузнецова Ю. А. Метод визуализации управляющих алгоритмов реального времени // Радіоелектронні і комп'ютерні системи. - 2012. - № 6. - С. 164–170.

[9] Мильман И. Е. Анализ данных о деятельности кредитных организаций с использованием программы интерактивного визуального анализа многомерных данных / И. Е.Мильман [и др.] // Научная визуализация. – 2015. – том 7. – №1. – С. 45-64.

[10] Пилюгин В. В. Научная визуализация как метод анализа научных данных / В. В. Пилюгин [и др.] // Научная визуализация. – 2012. – том 4. – №4. – С. 56-70.

[11] Хачумов В. М. , Виноградов А. Н.Разработка новых методов непрерывной идентификации и прогнозирования состояния динамических объектов на основе интеллектуального анализа данных // Матем. методы распознав. образов: 13-я Всероссиийская конференция. Ленинградская обл., Сборник докладов. – 2007. – С. 548-550.

[12] Шаропин К. А., Берестнева О. Г., Шкатова Г. И. Визуализация результатов экспериментальных исследований // Известия Томского политехнического университета. – 2010. – том 316. – №5. – С. 172-176.