Разработка и развитие системы интерактивного визуального анализа многомерных данных*

***О.П. Масленников*[†], *И.Е. Мильман*[†], *А.Э. Сафиуллин*[†], *А.Е. Бондарев*[‡], *Ш.У. Низаметдинов*[†], *В.В. Пилюгин*[†]**

maslolpavl@gmail.com|igalush@gmail.com|amir147@rambler.ru|bond@keldysh.ru
sh_nizam@mail.ru|VVPilyugin@mephi.ru

[†]НИЯУ МИФИ, Москва, Россия;
[‡]Институт прикладной математики им. М.В.Келдыша РАН, Москва, Россия

Работа посвящена развитию интерактивной системы для решения задач анализа многомерных данных интерактивным визуальным методом. Визуальная аналитика предоставляет удобные средства для решения задач анализа многомерных данных и в данной работе показано развитие инструментария для решения задачи кластеризации и дискриминантного анализа. Разрабатываемая система позволяет отобразить многомерное облако данных и проводить его анализ в пространствах меньшей размерности (2D и 3D), выдвигать и проверять различные гипотезы об исходных данных, с возможностью последующего построения предположений для проведения некоторых счетных методов, с помощью геометрических построений в интерактивном режиме.

Ключевые слова: *визуальная аналитика, анализ многомерных данных, интерактивный интерфейс*

Design and development of system for interactive visual analyzing of multidimensional data*

***O.P. Maslennikov*[†], *I.E. Milman*[†], *A.E. Safiulin*[†], *A.E. Bondarev*[‡], *Sh.U. Nizametdinov*[†], *V.V. Pilyugin*[†]**

[†]National Research Nuclear University MEPhI, Moscow, Russian Federation;
[‡]Keldysh Institute for Applied Mathematics RAS, Moscow, Russian Federation

The article presents a development of interactive system intended for visual analyzing of multidimensional data. The examples and illustrations are enclosed. Considered problems can be referred to visual analytics. By means of described interactive system user can work directly with data volume in question projections to 2D and 3D subspaces. Also the user is able to verify hypotheses about types of data inside the volume by interactive geometrical constructions.

Keywords: *visual analytics, multidimensional data, interactive system*

1. Введение

Современные задачи обработки и анализа огромных разнородных объемов информации требуют интенсивного развития методов, принципов и программных средств, позволяющих осуществить их решение. В роли средства решения выступает сравнительно молодая междисциплинарная ветвь исследований – визуальная аналитика. Методы визуальной аналитики интенсивно внедряются во все значимые прикладные аспекты человеческой деятельности. В практической плоскости визуальную аналитику можно рассматривать как решение задач анализа данных с использованием способствующего интерактивного визуального интерфейса, т.е. визуальная аналитика призвана организовать человеко-машинный интерфейс, усиливающий человеческие аналитические способности [1].

Основные методы, подходы и алгоритмы визуальной аналитики описаны в работах [2–6]. В этих же работах приведен ряд примеров современного применения визуальной аналитики в различных сферах человеческой деятельности, а также приведены описания ряда программных продуктов, построенных на основе визуальной аналитики [1].

Внимательное изучение литературы, посвященной описанию конкретных приложений в области визуальной аналитики, позволяет утверждать, что в реальности интерактивным системам работы с многомерными данными зачастую придается меньшее значение по сравнению с системами постпроцессинга результатов применения методов Data Analysis [1].

Данная работа представляет расширение разрабатываемой интерактивной системы визуального анализа многомерных данных. В рамках данной системы рассматриваются классические задачи анализа многомерных данных, такие как: выявление кластеров в многомерном облаке данных, построение системы решающих правил для процедур классификации объектов, реализация отображения многомерного объема данных в двумерных проекциях на все возможные пары координат. Разрабатываемая система позволяет пользователю:

Работа опубликована при финансовой поддержке РФФИ, грант 15-07-20347

— непосредственно работать с отображениями данных в пространствах меньшей размерности – двумерных и трехмерных;
— выдвигать гипотезы о наличии кластеров и классов в облаке данных и проверять их непосредственно с помощью интерактивного геометрического моделирования;
— выделять аномальные объекты в зависимости от расстояния между объектами многомерного пространства;
— принимать решения о возможности построения решающих правил для задач классификации новых объектов;
— проводить непосредственный поиск кластеров по множеству двумерных проекций и визуальный анализ значимости координатных направлений с точки зрения вклада в дисперсию.

Следует также отметить, что разрабатываемая интерактивная система, дает в перспективе возможность при дальнейшем применении математических методов анализа многомерных данных использовать полученные геометрические построения и гипотезы в качестве начальных приближений для более точных вычислений. При разработке интерактивной системы использовались материалы работ [7-10].

2. Интерактивный визуальный анализ

В рамках разрабатываемой интерактивной системы визуального анализа на сегодняшний день реализовано решение следующих задач.

Решение задачи кластерного анализа 3D проекционным методом

Решение данной задачи обеспечивает пользователю возможность интерактивной работы с проекциями исходного многомерного пространства в трехмерных пространствах, образованных из исходных координат по выбору пользователя. Пользователю предоставляется возможность интерактивного построения кластеров и выделение удаленных точек (аномальных объектов).

Идея метода заключается в том, что при проекции задается параметр d, отвечающий наибольшему внутрикластерному расстоянию. Если расстояние в исходном пространстве между двумя точками меньше чем d, то между данными точками строится отрезок. Точки во время проекции представляются сферами, а отрезки цилиндрами (Рис.1) [1].

Оптическая модель сцены предполагает присвоение цилиндрам цвета, отвечающему расстоянию между точками. Чем ближе расстояние к d, тем синее цилиндр.

Пользователю предлагается проводить анализ разбиения на подмножества в зависимости от параметра d.

Предусматривается два метода перехода – последовательный просмотр при задаваемых пользователем d или задание двух значений параметра и построение видеоряда.

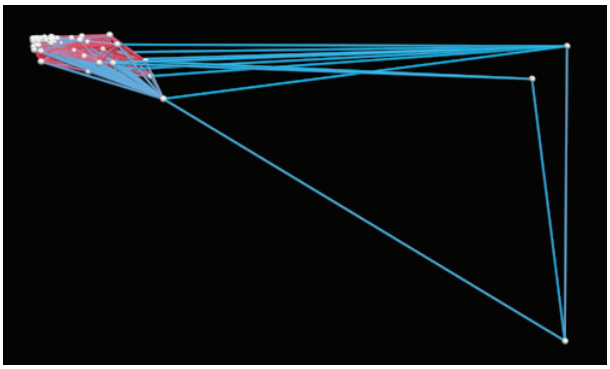

Рис. 1: Отображение множества точек

Возможными подмножествами являются:

Кластер – подмножество, при заданном множестве точек, попарное расстояние между которыми не превышает заданное d, а расстояние между точками кластера и остальными точками не меньше заданного d.
Удаленная (одиночная) точка – точка, удаленная от всех остальных точек исходного множества на расстояние, большее заданного d. В качестве примера можно привести зеленную точку на рис. 2.
Сгусток – подмножество точек, большая часть расстояний между которыми не превышает заданное d.
Квазиудаленная (Квазиодиночная) точка – точка, не являющаяся удаленной, но и не входящая в сгусток или кластер при заданном разбиении. В качестве примера можно привести желтую точку на рис. 2.

Определение точки к тому или иному подмножеству выполняется аналитиком в интерактивном режиме в зависимости от пространственной сцены, полученной при заданном d.

При разбиении на подмножества, пользователям предлагается использовать различные цветовые обозначения для распределения точек по подмножествам.

После получения удаленных точек, выполняется этап микроанализа. На данном этапе выявляется, какие координаты точек вносят наибольший вклад в расстояние между точками, а какие координаты влияют слабо.

Для этого предлагается строить графическую проекцию на плоскость (X_i, Y_i). При этом используются те же цветовые обозначения, введенные на предыдущем шаге.

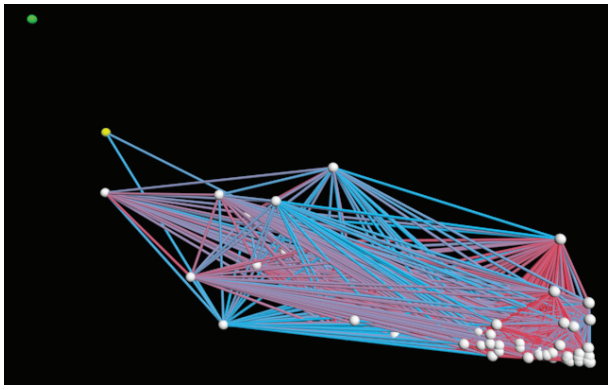

Рис. 2: Примеры подмножеств точек

На рис. 3 представлена плоская проекция результатов выполнения макроанализа на одну из координат. На данной проекции видно, что анализируемая координата вносит значительный вклад в удаленность зеленной точки, однако вносит малый в удаленность фиолетовой и розовой точек.

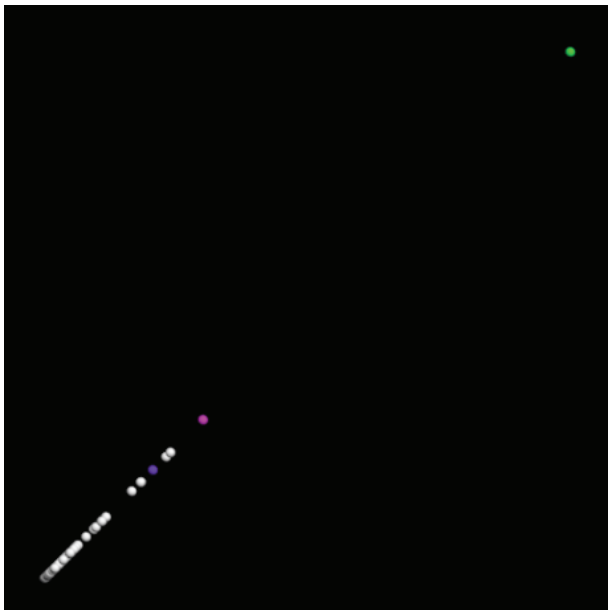

Рис. 3: Плоская проекция исходного подмножества

Решение задачи дискриминантного анализа 2D и 3D проекционным методом

Второй задачей, реализованной на сегодняшний день в разрабатываемой системе интерактивного визуального анализа многомерного облака данных, является классическая задача дискриминантного анализа. Задача решается с помощью проекционного метода, реализованного для 2D и 3D случаев.

Основным предположением дискриминантного анализа является то, что существуют две или более группы, которые по некоторым переменным отличаются от других групп, причем такие переменные могут быть измерены по интервальной шкале либо по шкале отношений. Дискриминантный анализ помогает выявлять различия между группами и дает возможность классифицировать объекты по принципу максимального сходства [10].

В качестве алгоритма решения задачи было предложено поочередно применять следующие методы:

1. Метод построения разделяющей гиперплоскости с построение последовательных проекционных изображений.

Гиперплоскость называется разделяющей для 2х групп точек, если при подстановке координат точки в аналитическое описание этой гиперплоскости все точки одной группы будут иметь один знак, а все точки другой группы – противоположный. Если в проекции плоскость является разделяющей, то при переходе к пространству с размерностью на единицу больше, данная плоскость будет являться так же разделяющей. Таким образом, предлагается последовательный обход через возможные 2х и 3х-мерные пространства и поиск в каждом из таких пространств разделяющей прямой или плоскости.

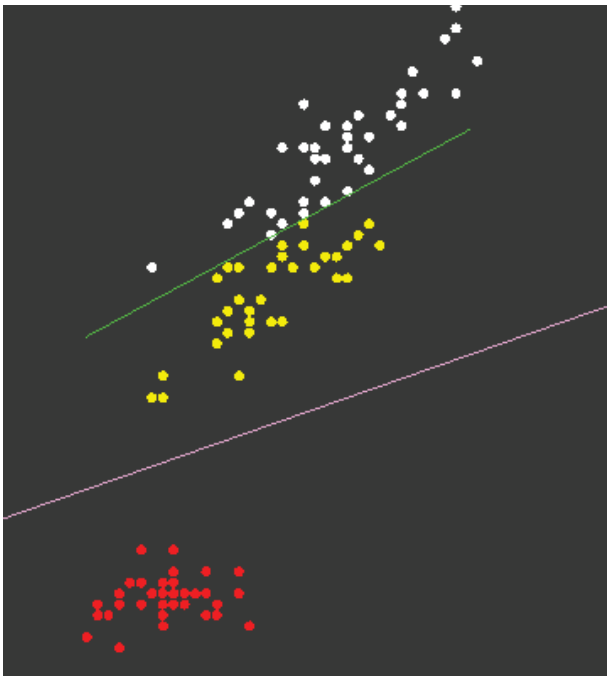

Рис. 4: Построение решающего правила в двумерном режиме работы

2. Переход в пространство главных компонент и применение метода построения разделяющей гиперплоскости в них.

В пространство главных компонент приводит линейное преобразование исходного пространства. При линейном преобразовании разделяющая плоскость продолжает являться разделяющей. Обратное преобразование также является линейным. Та-

ким образом, при построении решающего привала в пространстве главных компонент пользователь получает решающее правило и в исходном пространстве также. Переход в пространство главных компонент позволяет получить дополнительное пространство для поиска решения задачи и увеличение числа значимых коэффициентов в аналитическом описании гиперплоскости.

3. Метод построения решающего правила в виде теоретико-множественных операций.
В качестве решающего правила предложено строить систему гиперплоскостей. На двумерной проекции система гиперплоскостей представляет собой многоугольник, построенный пользователем. С помощью такой сложной фигуры можно разделить группы точек, которые не удалось разделись с помощью применения предыдущих методов. В качестве решающего правила будет выступать теоретико-множественное отношение в базисе «и, или, не», в качестве элементов множества выступают ребра многоугольника.

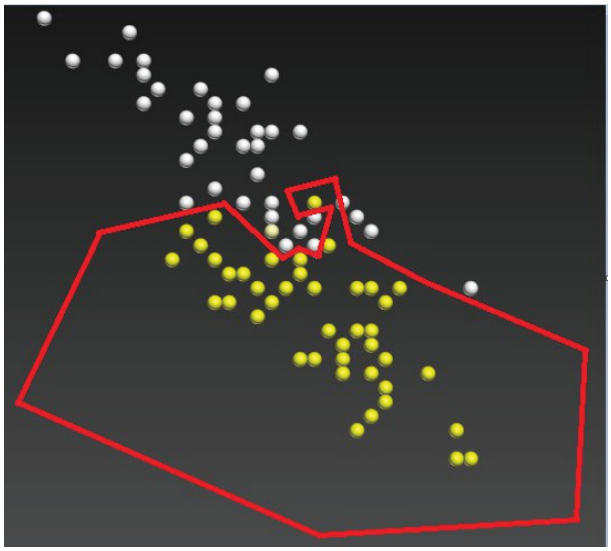

Рис. 5: Окаймляющий многоугольник и отношение, задающее его

После получения решающего правила решается задача о верификации новых объектов.

Решение задачи выделения кластеров 2D проекционным методом

В основе разрабатываемого метода разбиения исходного множества на кластеры лежит использование гипотезы компактности. Полагаясь на данную гипотезу, можно сделать суждение о том, что точки, близкие в каждом двумерном подпространстве, близки в n-мерном пространстве, а точки далекие в исходном пространстве, далеки в одном из двумерных подпространств.

Алгоритм решения задачи [11]:

1. Точки многомерного множества данных проецируются во все двумерные арифметические пространства; в итоге получается «матрица проекций».
2. На одной из проекций выделяются характерные образования – кандидаты на скопление.
3. Анализируются остальные проекции и в случае обнаружения точек, лежащих далеко от скопления, они исключаются.
4. Оставшиеся выделенные точки помечаются как кластер и исключаются из рассмотрения. Если сгустков не осталось (то есть остались лишь одиночные (изолированные) точки, либо все точки помечены как кластер) то задача решена, иначе следует вернуться к пункту 2.

Полученные визуальным методом кластеры при необходимости можно использовать как начальное приближение для модифицированного метода k-средних. Количество кластеров, подающихся на вход модифицированного алгоритма k–средних, равняется числу выделенных кластеров плюс число оставшихся одиночных точек. Для наглядности представления результатов кластеризации, помимо представления результатов в табличном виде (объект - номер кластера), было решено реализовать метод двумерного представления объектов кластеризации, называемый профильной диаграммой (диаграмма с параллельными координатами).

После проведения кластеризации, объекты, объединенные в одну группу, помечаются соответствующим номеру кластера цветом. Эта возможность позволяет оценивать адекватность полученного разбиения и помогает в его интерпретации. Представив таким образом полученные результаты, пользователь может определить, какие объекты являются спорными, т.е. могут принадлежать как одному, так и другому кластеру, какие группы вообще не имеют места быть, в силу невозможности их интерпретации с точки зрения здравого смысла.

Так же в качестве одного из методов визуализации иерархической кластеризации применяются дендрограммы. В случаях, когда число признаков велико (больше 40) диаграмму рассеяния становится сложно визуально анализировать. С помощью метода главных компонент программа позволяет снизить размерность признакового пространства. На рис. 1. представлены многомерные данные – изображения лиц двух национальностей. После сокращения размерности представляется возможность наблюдать диаграммы рассеяний – проекции изображений в признаковое подпространство. Как видно по первым двум компонентам отчетливо видна разделимость двух групп.

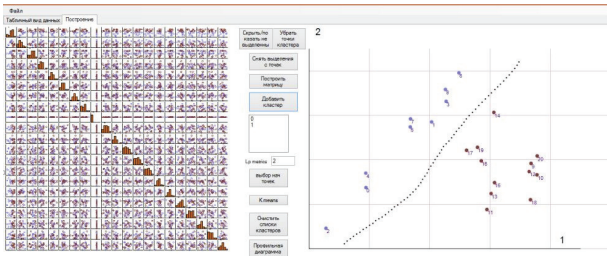

Рис. 6: Диаграммы рассеяния и увеличенная проекция значений на двумерное подпространство из первых двух главных компонент

3. Заключение

В данном докладе представлен ряд реализованных задач, относящихся к развивающейся в настоящее время интерактивной системе визуального анализа многомерных данных. Основная цель данной разработки – предоставление пользователю возможности интерактивной работы с двумерными и трехмерными проекциями исходного многомерного объема данных для получения первичной информации о структуре изучаемого объема и взаимном расположении точек в изучаемом объеме.

Литература

[1] Масленников О.П., Мильман И.Е., Сафиуллин А.Э., Бондарев А.Е., Низаметдинов Ш.У., Пилюгин В.В. Интерактивный визуальный анализ многомерных данных / ГрафиКон'2014: 24-я Международная конференция по компьютерной графике и зрению: Ростов-на-Дону, Академия архитектуры и искусств ЮФУ Труды конференции. – С.51–54.

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

[3] Keim D. A, Mansmann F, Schneidewind J, Thomas J, Ziegler H: Visual analytics: Scope and challenges, Visual Data Mining: 2008, S. 82.

[4] Keim D., Andrienko G., Fekete J.-D., Gorg C., Kohlhammer J., and Melancon G. "Visual Analytics: Definition, Process, and Challenges", A. Kerren et al. (Eds.): Information Visualization, LNCS 4950, pp. 154–175, 2008. Springer-Verlag Berlin Heidelberg 2008.

[5] Kielman, J. and Thomas, J. (Guest Eds.) (2009). Special Issue: Foundations and Frontiers of Visual Analytics, Information Visualization, Volume 8, Number 4, Winter 2009, p. 239-314.

[6] Keim D., Kohlhammer J., Ellis G. and Mansmann F. (Eds.), Mastering the Information Age – Solving Problems with Visual Analytics, Eurographics Association, 2010.

[7] Пилюгин В.В., Маликова Е.Е., Пасько А.А., Аджиев В.Д. Научная визуализация как метод анализа научных данных / Научная визуализация. Т.4, № 4, с.8-25, 2012, URL: http://sv-journal.org/2012-4/06.php?lang=ru

[8] Бондарев А.Е., Галактионов В.А. Анализ многомерных данных в задачах многопараметрической оптимизации с применением методов визуализации / Научная визуализация. Т.4, № 2, с.1-13, 2012, URL: http://sv-journal.org/2012-2/01.php?lang=ru

[9] Основы научной визуализации [сайт]. URL: http://sv-journal.org/unl (дата обращения: 10.05.2015)

[10] Низаметдинов Ш.У. Анализ данных. М.: МИФИ, 2006

[11] Масленников О. П., Мильман И.Е., Сафиуллин А.Э., Бондарев А.Е., Низаметдинов Ш.У., Пилюгин В.В. Разработка системы интерактивного визуального анализа многомерных данных//Научная визуализация, МИФИ, 2014, 4, С.30-49. URL: http://sv-journal.org/2014-4/04.php?lang=ru