Интерактивный визуальный анализ многомерных данных

О. Масленников*, И. Мильман*, А.Сафиуллин*, А. Бондарев**, Ш. Низаметдинов*, В. Пилюгин*
НИЯУ МИФИ*
Институт прикладной математики им. М.В. Келдыша РАН**
Москва, Россия
maslolpavl@gmail.com; igalush@gmail.com; amir147@rambler.ru; bond@keldysh.ru; sh_nizam@mail.ru; pilyugin@sv-journal.com

Аннотация

Работа посвящена вопросам разработки интерактивной системы, предназначенной для решения задач визуального анализа многомерных данных. Рассматриваемые в примерах и иллюстрациях задачи относятся к области визуальной аналитики. Система позволяет непосредственную работу пользователя с визуальными представлениями многомерного облака данных в пространствах меньшей размерности, выдвижение и проверку гипотез о строении и характере изучаемых данных с помощью геометрических построений в интерактивном режиме.

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

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

Основные методы, подходы и алгоритмы визуальной аналитики описаны в работах [1 – 5]. 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.

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 отображения результатов применения методов Data Analysis.

Данная работа представляет разрабатываемую интерактивную систему визуального анализа многомерных данных. В рамках разрабатываемой системы рассматриваются классические задачи анализа многомерных данных, такие, как: построение кластеров и их оболочек 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:

- 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;

- строить оболочки обнаруженных кластеров, максимально приближенные к данным, в системе координат главных компонент;

- 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 [6 - 9].

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 их оболочек различными способами.

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).

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

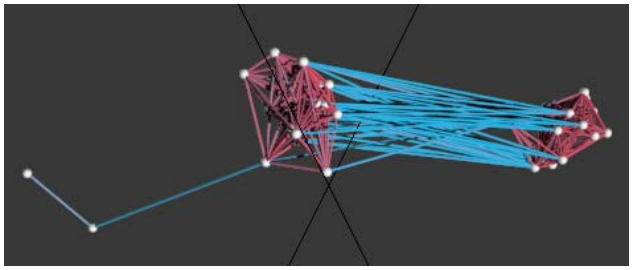

Рис. 1. Отображение множества точек

Пользователю предлагается проводить анализ разбиения на кластеры в зависимости от параметра d. Предусматривается два метода перехода – последовательный просмотр при задаваемых пользователем d или задание двух значений параметра и построение видеоряда.

Анализ формы кластера предлагается методом построения оболочки в трехмерном пространстве. Построение оболочки возможно несколькими методами – построение прямоугольного параллелепипеда методом пересечения сфер и смешанным методом.

В первом методе прямоугольный параллелепипед строится на осях, полученных с помощью метода главных компонент. Данное построение гарантирует хорошее приближение параллелепипеда к кластеру (рис. 2).

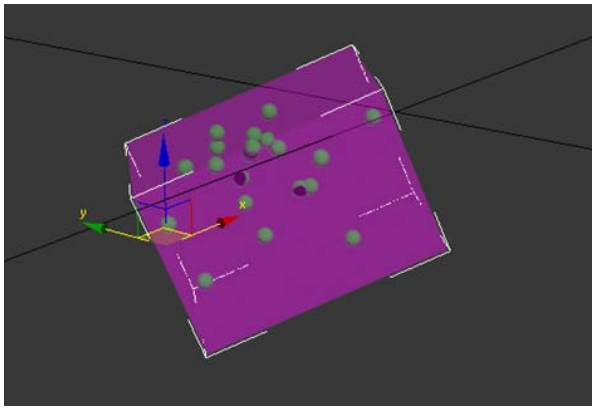

Рис. 2. Оболочка в виде прямоугольного параллелепипеда для кластера

Во втором методе строятся сферы с центром в каждой точке кластера и радиусом, равном максимальному расстоянию от этой точки до точек кластера (оно заведомо меньше либо равно d). А затем строится пересечение сфер. Полученная оболочка, очевидно, включает в себя все точки и дает достаточно хорошее приближение (рис.3). Данная оболочка лучше предыдущей при относительно одинаковых собственных значениях матрицы корреляций внутри кластера.

Смешанный метод построения оболочки предполагает построения оболочек по двум описанным методам и дальнейшее пересечение двух оболочек.

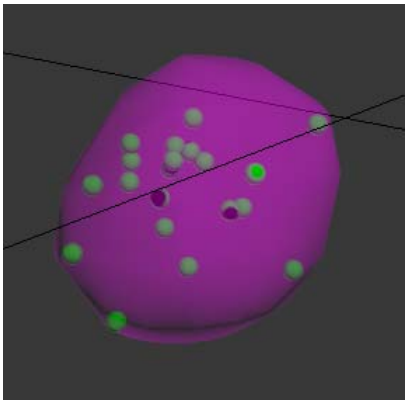

Рис. 3. Оболочка, построенная с помощью пересечения сфер

Решение задачи дискриминантного анализа 2D и 3D проекционным методом

Решение данной задачи предоставляет пользователю возможность построения плоскости, разделяющей классы точек, при работе с проекциями исходного многомерного объема данных на двумерные и трехмерные подпространства, образованные из исходных координат.

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.

Основным методом решения задачи дискриминантного анализа является метод нахождения коэффициентов гиперплоскости Фишера [9]. В результате исследования методов решения задачи в качестве метода решения был предложен метод построения разделяющей гиперплоскости при помощи метода последовательных проекционных изображений. Суть метода заключается в том, что если мы можем в проекции построить разделяющую плоскость, то при переходе к пространству с размерностью на единицу больше, данная плоскость будет являться также разделяющей. В качестве алгоритма решения был предложен последовательный просмотр и анализ 2х и 3х-мерных проекций с целью нахождения разделяющей плоскости или системы плоскостей (рис. 4).

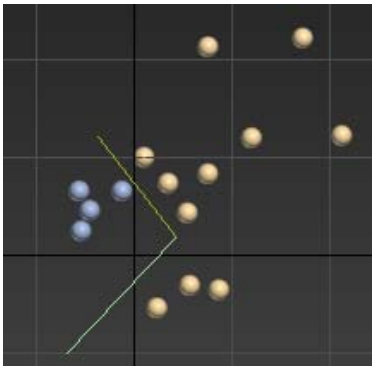

Рис. 4. Система разделяющих прямых для двух групп точек

После нахождения такой системы пользователю предоставляется возможность проведения верификации построенной формальной системы решающих правил. После проведения верификации пользователь может решать задачу классификации новых объектов, добавляемых к исходному многомерному объему данных (рис. 5).

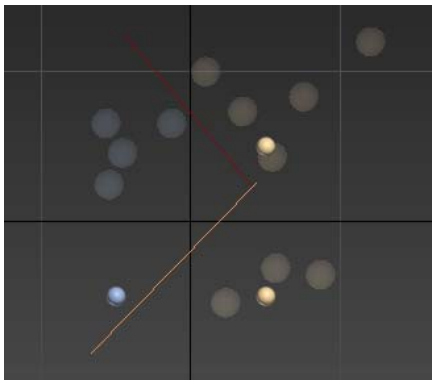

Рис. 5. Решение задачи классификации новых объектов

Данный метод позволяет принимать решения о возможности построения решающих правил для задач классификации новых объектов.

Решение задачи выделения кластеров 2D проекционным методом

Цель решения данной задачи в разрабатываемой интерактивной системе – предоставление пользователю возможности одновременной работы со всеми проекциями исходного многомерного объема на двумерные подпространства, образованные из исходных координат. Исходя из соображения, что точки, близкие во всех двумерных проекциях, будут близки и в исходном многомерном пространстве, пользователь интерактивной системы может выделять близкие точки на двумерной проекции, маркировать их цветом, вносить или удалять точки. Все действия пользователя отображаются одновременно на всех двумерных проекциях.

Алгоритм решения задачи:

Этап 1. Точки исходного многомерного пространства проецируются на двумерные подпространства, образованные из исходных координат. Таким образом, получается матрица проекций (рис. 6).

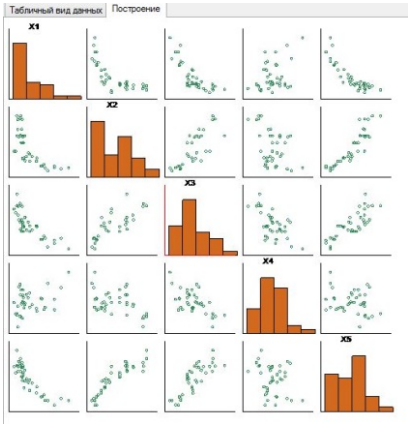

Рис. 6. Матрица проекций

Этап 2. На одной из проекций выделяются характерные образования – кандидаты на сгустки.

Этап 3. Проводится анализ всех остальных проекций, при обнаружении точек, лежащих далеко от сгустка, данные точки исключаются из рассмотрения.

Этап 4. Оставшиеся выделенные точки помечаются как кластер и исключаются из дальнейшего рассмотрения. Если не осталось сгустков точек (одиночные точки, либо все точки помечены как кластер), то переходим к этапу 5, иначе – к этапу 2.

Этап 5. В результате пользователем получена первичная картина кластерного разбиения изучаемого многомерного объема данных. Для дальнейшего улучшения картины в системе реализован алгоритм к-средних.

Для наглядности представления результатов кластеризации в системе реализовано построение профильной диаграммы (рис. 7) как способа двумерного представления объектов кластеризации.

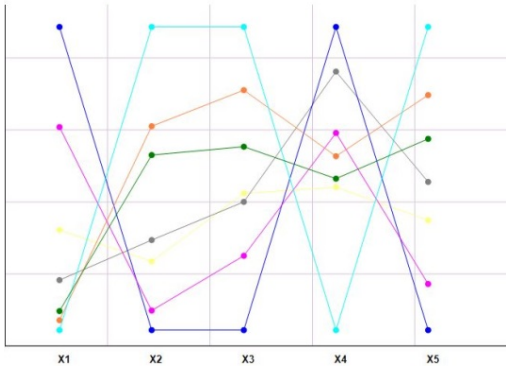

Рис. 7. Профильная диаграмма

В результате работы с интерактивной системой пользователь может также определить, какие объекты являются спорными, т.е. могут принадлежать одновременно к нескольким кластерам.

3. ЗАКЛЮЧЕНИЕ

В данном докладе представлен ряд реализованных задач, относящихся к разрабатываемой 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.

4. ССЫЛКИ

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

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

[3] 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. – P. 154 – 175, 2008. Springer-Verlag Berlin Heidelberg 2008.

[4] 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.

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

[6] 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.com/2012-4/06.php?lang=ru*

[7] 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.com/2012-2/01.php?lang=*ru

[8] *Основы научной визуализации [сайт]. URL: http://edu-cons.net/unl/ (дата обращения: 10.05.2014)*

[9] Низаметдинов Ш.У. *Анализ данных. – М.: МИФИ, 2006.*

INTERACTIVE VISUAL ANALYZING OF MULTIDIMENSIONAL DATA

Abstract

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 his hypotheses about types of data inside the volume by interactive geometrical constructions.

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

Об авторах

Масленников Олег – студент НИЯУ МИФИ.

E-mail: maslolpavl@gmail.com

Сафиуллин Амир – студент НИЯУ МИФИ.

E-mail: amir147@rambler.ru

Мильман Игаль – студент НИЯУ МИФИ. E-mail: igalush@gmail.com

Бондарев Александр – к.ф.-м. н., старший научный сотрудник ИПМ им. М.В. Келдыша РАН.
E-mail: bond@keldysh.ru

Низаметдинов Шамиль – к.т.н., доцент кафедры системного анализа НИЯУ МИФИ.

E-mail: sh_nizam@mail.ru

Пилюгин Виктор – к.т.н., профессор, заведующий лабораторией «Научная визуализация» НИЯУ МИФИ.

E-mail: pilyugin@sv-journal.com