Анализ и визуализация многомерных данных в задачах вычислительной газовой динамики

А. Бондарев, В. Галактионов, Л. Шапиро
Институт прикладной математики им. М.В. Келдыша РАН
Москва, Россия
bond@keldysh.ru; vlgal@gin.keldysh.ru; pls@gin.keldysh.ru

Аннотация

В докладе рассматриваются вопросы анализа и визуализации многомерных объемов данных в задачах вычислительной газовой динамики. Многомерные данные появляются в вычислительной газовой динамике как результаты параметрических исследований и решения задач оптимизационного анализа. Предлагается приближенный подход, предполагающий отображение многомерного объема в пространстве главных компонент и аппроксимацию данных в объеме с помощью плоскостей. Приведен пример практического применения подхода.

Ключевые слова: *визуализация многомерных данных, метод главных компонент, вычислительная газовая динамика*

1. ВВЕДЕНИЕ

Задачи обработки, анализа и визуализации многомерных данных являются на сегодняшний день важным и актуальным направлением. Проблемы изучения многомерных объемов данных, задачи определения взаимного расположения точек в многомерном облаке данных, задачи выявления определяющих факторов и скрытых взаимосвязей между ними возникают практически во всех областях знания. Анализ многомерных данных (Data Analysis) интенсивно развивается как научная дисциплина, которая включает в себя: метод главных компонент (PCA-Principal Component Analysis) и его обобщения на нелинейные случаи, факторный анализ, кластерный анализ, дискриминантный анализ, построение самоорганизующихся карт (SOM – Self-Organized Maps) и упругих карт (Elastic Maps) [1,2]. Комбинированное применение методов, алгоритмов и подходов, разработанных в этих разделах, позволяет провести всестороннее исследование многомерного объема данных вне зависимости от их происхождения.

В задачах вычислительной газовой динамики проблемы анализа многомерных данных ранее практически не встречались. Для обработки и визуального представления результатов даже самых сложных расчетов вполне хватало наработанных методов и приемов научной визуализации [3]. Однако в настоящее время интенсивное развитие высокопроизводительных и параллельных вычислений позволяет решать задачи параметрического исследования и задачи оптимизационного анализа [4, 5].

Параметрические численные исследования позволяют получать решение не для одной конкретной задачи математического моделирования, а для класса задач, заданного в многомерном пространстве определяющих параметров. Также применение параллельных алгоритмов на высокопроизводительной вычислительной технике позволяет численное исследование задач оптимизационного анализа, когда обратная задача решается в каждой точке сеточного разбиения многомерного пространства определяющих параметров. Основная особенность, с точки зрения задач анализа и визуализации решений, в подобных вычислениях заключается в том, что их результаты представляют собой многомерные массивы, размерность которых соответствует количеству определяющих параметров. Эти массивы нуждаются в обработке и визуальном представлении с целью их анализа и выявления внутренних взаимосвязей между определяющими параметрами. Подобные задачи начинают встречаться на практике все чаще, хотя следует отметить, что размерность подобных массивов сегодня ограничивается вычислительными мощностями и обычно составляет 4 – 5, в исключительных случаях – 6.

В этой ситуации естественно хотелось бы применить уже наработанный аппарат методов и алгоритмов Data Analysis к подобным задачам. Однако здесь возникают некоторые проблемы, обусловленные спецификой целей исследования и происхождения самих данных. В задачах Data Analysis многомерные данные рассматриваются как набор точек $A_i(x_1,...,x_n), i = 1,...,m$, и основной интерес для исследователя представляет их взаиморасположение с целью выделения кластеров, решения задачи классификации новых объектов. Когда мы рассматриваем многомерные данные в задачах вычислительной газовой динамики (CFD), полученные как результаты решения задач оптимизационного анализа или параметрических исследований, нас в гораздо меньшей степени интересует взаиморасположение точек, так как разбиения по определяющим параметрам $x_1,...,x_{n-1}$ задаются нами. Основная цель здесь – изучение зависимости $x_n = F(x_1,...,x_{n-1})$, представленной по результатам вычислений в виде многомерного объема данных, визуализация этой зависимости и, по возможности, представление ее в квазианалитическом виде с помощью приближений.

Таким образом, задача адаптации методов Data Analysis для целей исследования многомерных результатов расчетов газодинамических задач является актуальной.

2. АНАЛИЗ И ВИЗУАЛИЗАЦИЯ МНОГОМЕРНЫХ ДАННЫХ

Наиболее эффективным путем анализа многомерных данных, получаемых в результате решения задач вычислительной газовой динамики, является визуальное представление зависимости $x_n = F(x_1,...,x_{n-1})$ и получение информации о характере этой зависимости. Далее следует аппроксимация

зависимости с помощью поверхностей достаточно простого вида и получение, как следствие, искомого квазианалитического выражения.

В работе [4] рассматривались современные попытки построения визуальной концепции для представления многомерных данных, а также отмечалось отсутствие в настоящее время адекватного и надежного способа подобного визуального представления для объемов, имеющих размерность, превышающую 3. Следовательно, для анализа информации, содержащейся в полученном многомерном массиве, необходимо понизить размерность массива. Рассмотрим наиболее распространенные практические способы понижения размерности.

Рассматриваемые способы основаны на анализе дисперсий данных массива по координатным направлениям или нахождении в изучаемом многомерном пространстве вектора, по направлению которого дисперсия максимальна.

Первый способ представляет собой поиск координатного направления с наименьшей дисперсией. Вычисляются дисперсии D_i по всем координатным направлениям, выбирается наименьшая из них, и в том случае, когда минимальная дисперсия существенно (на порядки) меньше остальных, значения исследуемой функции по координатному направлению с наименьшей дисперсией заменяются на константу, равную среднему значению по направлению. Таким образом, размерность исходного многомерного пространства понижается на единицу.

Более радикальный вариант данного способа выглядит следующим образом. Вычисляются дисперсии по всем координатным направлениям и ранжируются в порядке убывания. Выбираются три направления, соответствующих максимальным дисперсиям D'_1, D'_2, D'_3. Далее проверяется условие $D'_j >> \varepsilon * D'_i, i \neq 1,2,3$, где ε – малая величина, $j=1,2,3$

задаваемая пользователем. Если это условие выполнено, то полагаем значения искомой функции по всем направлениям, кроме трех, соответствующих максимальным дисперсиям, константами, равными соответствующим средним значениям по направлениям. Таким образом, мы радикально понижаем размерность исходного пространства и оказываемся в рамках стандартного трехмерного пространства.

Изложенный подход обладает рядом недостатков:

- он далеко не всегда осуществим, например, если данные в многомерном пространстве близки к гиперсфере;

- в выборе малой величины ε заложен произвол.

Однако, несмотря на эти недостатки, для пространств небольшой размерности $n = 4,5$ во многих практических случаях данных подход работает успешно.

Второй распространенный способ понижения размерности заключается в построении графических проекций на стандартное число измерений $n \leq 3$ с фиксацией переменных, не участвующих в построении проекции. В тех случаях, когда из набора дисперсий по направлениям нельзя выделить существенно наименьшую, часто используется метод разделения переменных.

Если из вида проекций в стандартных измерениях удается сделать вывод о том, что для двух переменных при фиксированных остальных переменных исследуемая функция может быть выражена с помощью аналитической зависимости Φ_1, а для остальных переменных при фиксированных первых двух – с помощью зависимости Φ_2, то выдвигается гипотеза о том, что итоговая аналитическая зависимость для искомой функции может быть представлена в виде комбинации этих функций со сшивкой при фиксированных значениях.

Оба вышеизложенных подхода не являются строго обоснованными. Скорее, это алгоритмы выдвижения гипотез, нуждающихся в проверках. Однако эти методы позволяют получать реальные практические результаты.

Не менее эффективным является применение метода главных компонент (PCA). Суть метода состоит в переходе от исходной системы координат к новому ортогональному базису в рассматриваемом многомерном пространстве, оси которого ориентированы по направлениям максимальной дисперсии массива данных. Реализации метода главных компонент и алгоритмам его применения в различных областях посвящено большое количество литературы. Различные варианты реализации метода главных компонент и его обобщений для нелинейных случаев подробно представлены в работах [1, 2]. Геометрическая постановка задачи нахождения главных компонент формулируется согласно [1, 2] следующим образом. В многомерном пространстве ищется вектор направления, задающий прямую, вдоль которой дисперсия максимальна (или сумма квадратов расстояний от точек данных до прямой минимальна). Так определяется первая главная компонента. Далее рассчитывается множество векторов первых остатков, которое лежит в пространстве, ортогональном первой главной компоненте и имеющем размерность на единицу меньше исходной размерности. Для нового пространства, образованного этим множеством векторов, снова ищется направление с максимальной дисперсией. Так рассчитывается вторая главная компонента. Снова рассчитывается множество векторов вторых остатков и т.д.

Применение главных компонент дает нам возможность отобразить исследуемый многомерный массив на плоскость или в трехмерное пространство, образованное первыми тремя главными компонентами. В этом случае схема обработки, анализа и визуализации исходного многомерного объема данных будет выглядеть следующим образом.

- Для исходного объема вычисляются три первые главные компоненты Y_1, Y_2, Y_3, где каждая главная компонента является линейной комбинацией исходных переменных $Y(x_1,...x_n) = \sum B_i x_i$.

- Далее координаты исходных точек исследуемого объема выражаются в координатах главных компонент:

$$A_i(x_1,...x_n) = A_i(Y_1(x_1,...x_n), Y_2(x_1,...x_n), Y_3(x_1,...x_n))$$ -

Реализуется визуальное представление массива в двумерном виде $A_i(Y_1, Y_2)$ или в трехмерном $A_i(Y_1, Y_2, Y_3)$.

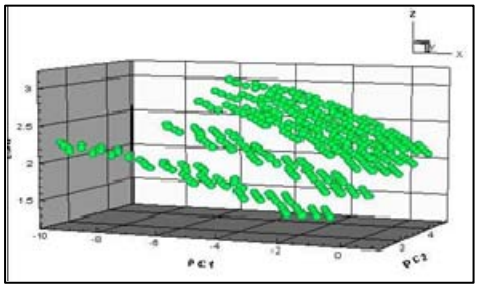
- Далее изучается полученное визуальное представление многомерного массива в главных компонентах и предпринимается попытка аппроксимации данных массива с помощью примитивных функций, имеющих аналитическое выражение. В простейшем случае применяется грубая

линейная аппроксимация с помощью плоскости вида $E_1Y_1 + E_2Y_2 + E_3Y_3 = C_y$. Так как плоскость при переходе от главных компонент к исходным переменным сохраняет свои свойства, с помощью обратного преобразования получаем $E_1^{'}x_1 + E_2^{'}x_2 + ... + E_n^{'}x_n = C_x$, которое уже можно рассматривать как искомую квазианалитическую зависимость $x_n = F(x_1,...,x_{n-1})$. В том случае, когда $A_i(Y_1, Y_2, Y_3)$ нельзя аппроксимировать одной плоскостью, можно использовать кусочно-линейный подход, применив несколько плоскостей. Следует также заметить, что применение квадратичных поверхностей может также оказаться весьма полезным, однако этот вопрос заслуживает отдельного рассмотрения.

3. ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ

Данный приближенный подход был применен к многомерному объему данных, полученному как результат решения задачи оптимизационного анализа нестационарного взаимодействия сверхзвукового потока вязкого сжимаемого теплопроводного газа со струйной преградой [5]. Преграда возникает благодаря недорасширенной спутной струе, истекающей из сопла, помещенного во внешний сверхзвуковой поток. При повышении скорости изменения степени нерасчетности струи возникает специфический режим течения, когда вещество струи распространяется вверх по потоку, по внешней стенке сопла. Скорость изменения степени нерасчетности струи рассматривается как управляющий параметр задачи оптимизационного анализа. В качестве определяющих параметров задачи рассматривались характерные числа Маха, Рейнольдса, Прандтля и Струхаля. Эти четыре параметра варьировались в определенных диапазонах. Целью решения задачи было нахождение скорости изменения степени нерасчетности струи, при которой реализуется специфический режим течения во всех диапазонах изменения характерных чисел задачи.

В качестве результата решения задачи был получен 5-мерный объем данных, где в качестве переменных были 4 характерных числа задачи $M_\infty, \lg \mathrm{Re}_\infty, \mathrm{Pr}, Sh_\infty$ и искомая скорость V^*. Для полученного многомерного объема были определены три первые главные компоненты. После перехода к главным компонентам строилось визуальное представление точек массива в главных компонентах (см. рисунок). Полученное визуальное представление многомерного массива в главных компонентах позволило предположить, что точки массива могут быть грубо аппроксимированы параметрически заданной плоскостью.



Представление многомерного объема в пространстве трех первых главных компонент

После определения конкретного вида плоскости и ее коэффициентов было проведено обратное преобразование к исходным переменным и определение конкретного вида аппроксимирующей плоскости в исходных координатах. Это дало возможность получить искомую зависимость $V^* = F(M_\infty, \lg \mathrm{Re}_\infty, \mathrm{Pr}, Sh_\infty)$ в аналитическом виде. Полученные результаты представляют собой решение для класса задач, заданного в многомерном объеме определяющих параметров.

4. ЗАКЛЮЧЕНИЕ

Предложенный приближенный подход, предполагающий отображение многомерного объема в пространстве главных компонент и аппроксимацию данных в объеме с помощью плоскостей, позволяет проводить оценку скрытых взаимозависимостей в многомерных объемах данных, получаемых в задачах вычислительной газовой динамики, как результаты решения задач параметрического поиска и оптимизационного анализа.

5. БЛАГОДАРНОСТИ

Данная работа выполнена при поддержке гранта Российского фонда фундаментальных исследований (проект N 14-01-00769а)

6. ССЫЛКИ

[1] 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.

[2] Зиновьев А. Ю. Визуализация многомерных данных, Красноярск, Изд. КГТУ, 2000. – 180 с.

[3] Бондарев А.Е., Галактионов В.А., Чечеткин В.М. Анализ развития концепций и методов визуального представления данных в задачах вычислительной физики // Вычислительная математика и математическая физика, 2011. – Т. 51, N 4. – С. 669 – 683.

[4] Бондарев А.Е., Галактионов В.А. Анализ многомерных данных в задачах многопараметрической оптимизации с применением методов визуализации // Научная визуализация, 2012. – Т. 4, № 2. – С. 1 - 13,

[5] Bondarev A.E, Galaktionov V.A. Parametric Optimizing Analysis of Unsteady Structures and Visualization of Multidimensional Data // International Journal of Modeling, Simulation and Scientific Computing, Vol. 4, suppl. issue 1, 2013, DOI: 10.1142/S1793962313410043 http://www.worldscientific.com/doi/abs/10.1142/S1793962313410043

Abstract

This article is devoted to the questions of multidimensional data analysis and visualization for CFD problems. Multidimensional data are considered as results of parametrical search and optimizing analysis. Rough approximate approach is proposed for data processing. The approach includes data visualization in principal components and data approximation by planes. An example of approach application to practical problem is considered.

Keywords: *multidimensional data visualization, PCA, CFD problems*

Об авторах

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

Галактионов Владимир – д.ф.-м. н., профессор, заведующий отделом компьютерной графики ИПМ им. М.В. Келдыша РАН.
E-mail: vlgal@gin.keldysh.ru

Шапиро Лев З – к.т.н., старший научный сотрудник ИПМ им. М.В. Келдыша РАН.
E-mail: pls@gin.keldysh.ru