Автоматическое приближение набора точек поверхностями второго порядка в задачах анализа многомерных данных*

Д.А. Сумин, А.Е. Бондарев, А.Г. Волобой
da.sumin@gmail.com|bond@keldysh.ru|voloboy@gin.keldysh.ru
Институт прикладной математики им. М.В. Келдыша РАН Москва, Россия

В статье рассматривается задача выявления закономерностей между переменными многомерных объёмов данных. Используется подход, основанный на понижении размерности исходных многомерных данных, позволяющий сводить исходную задачу к анализу набора точек в трёхмерном пространстве. Предлагается автоматическая аппроксимация трёхмерного набора данных семейством поверхностей второго порядка; в случае нахождения таких поверхностей становится возможным сделать выводы о взаимосвязи соответствующих исходных переменных.
Рассмотрен метод обработки набора точек в трёхмерном пространстве, состоящий из разделения набора точек на поднаборы и последующего определения приближающих поверхностей второго порядка. Исследованы и приведены ограничения, которые рассматриваемый подход накладывает на входные данные.

Ключевые слова: *обработка многомерных данных, метод главных компонент, сегментация облаков точек, методы подгонки.*

Automatic fitting of 3D point clouds with quadric surfaces in multidimensional data analysis*

D.A. Sumin, A.E. Bondarev, A.G. Voloboy
Keldysh Institute of Applied Mathematics, RAS, Moscow, Russian Federation

In this article, we address a problem of finding relationships between variables of multidimensional datasets. A PCA-based approach is used for reducing problem's dimension to perform processing in a 3D space. An automated fitting of 3D-data to quadric surfaces is used to reveal hidden relationships between variables of original multidimensional data.
Proposed approach of multidimensional data processing bases on dimensional-reduction methods and 3D-data processing algorithms: segmentation of point clouds and fitting of point clouds to quadric surfaces. The latter methods are reviewed and examined.

Keywords: *multidimensional-data processing, PCA, point-cloud segmentation, fitting methods.*

Введение

Необходимость работы с многомерными данными возникает сегодня в большом количестве прикладных и теоретических областей науки. Актуальными являются задачи анализа и визуализации многомерных данных, задачи определения взаимного расположения точек в многомерных облаках данных. Интересными являются вопросы выявления определяющих факторов и скрытых взаимосвязей между ними.

Существует большое количество практических задач, в которых необходимо оптимизировать состояние системы, зависящей от большого (более двух) количества параметров. Задача в этом случае является многомерной, и к ней оказывается невозможным применить классические методы визуализации с целью определить влияние каждого из параметров и их комбинации на систему. Зачастую при этом проведение эксперимента с каждой новой комбинацией параметров может быть сопряжено с различными трудностями и затратами; перебор всех необходимых вариантов может оказаться очень сложным или невозможным. В этой ситуации было бы логично применить математический аппарат для предположения результатов экспериментов без их фактического проведения.

В [1] рассмотрен подход к анализу многомерных данных, возникающих в задачах вычислительной газовой динамики. Целью анализа набора точек $A_i(x_1, ..., x_n), i = 1, ..., m$ являлось изучение зависимости $x_n = F(x_1, ..., x_{n-1})$ и поиск её явного представления. Задачу в такой постановке можно рассмотреть достаточно широко и применить методы её решения к описанным выше прикладным задачам предсказания состояния многопараметрических систем и нахождения скрытых зависимостей между переменными многомерных данных, поэтому развитие методов решения таких задач представляется актуальным. В этой статье описано развитие предложенного в [1] подхода.

Описание базового подхода

В этом разделе будет кратко освещен подход, изложенный в [1]. Авторы утверждают, что визуальное представление исследуемой зависимости $x_n = F(x_1, ..., x_{n-1})$ — наиболее эффективный способ

Работа выполнена при финансовой поддержке РФФИ, грант 14-01-00769 и опубликована при финансовой поддержке РФФИ, грант 15-07-20347.

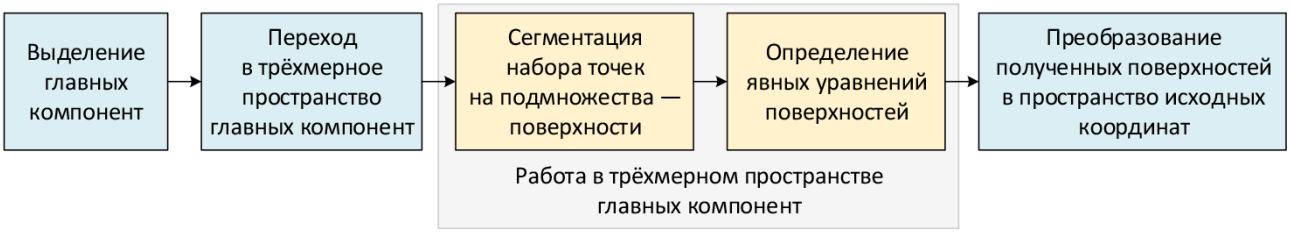

Рис. 1: Общая схема предлагаемого алгоритма.

анализа многомерных данных. Первым шагом обработки входных многомерных данных является понижение их размерности до трёх переменных, так как в настоящее время отсутствует [2] надежный и адекватный способ визуализации данных, имеющих размерность, превышающую 3.

Для понижения размерности входных данных $A_i(x_1, ..., x_n), i = 1, ..., m$ с целью дальнейшего поиска зависимости $x_n = F(x_1, ..., x_{n-1})$ предлагается применить метод главных компонент (PCA), который является одним из основных методов уменьшить размерность данных, потеряв при этом минимальное количество информации. Реализации метода главных компонент и обобщения на нелинейные случаи подробно описаны в [2]; применение метода главных компонент в визуализации данных описано в [3].

После применения метода главных компонент будут получены новые переменные Y_1, Y_2, Y_3, которые являются линейными комбинациями исходных переменных: $Y_j(x_1, ..., x_n) = \sum B_i x_i, j = 1, 2, 3$. Это позволяет выразить точки исходного набора данных в трёхмерном пространстве главных компонент: $A_i(x_1, ..., x_n) = A_i(Y_1(x_1, ..., x_n), Y_2(x_1, ..., x_n), Y_3(x_1, ..., x_n))$.

После этого предполагается визуализация набора точек в трёхмерном пространстве и её изучение на предмет возможности аппроксимировать набор точек с помощью функций, имеющих аналитическое выражение. В [1] авторы применяют аппроксимацию с помощью параметрически заданных плоскостей вида $E_1Y_1+E_2Y_2+E_3Y_3 = C_y$. Вследствие того, что плоскость при обратном переходе к исходным переменным сохраняет свои свойства, в результате преобразования получается следующая зависимость: $E'_1Y_1+E'_2Y_2+...+E'_nY_n = C_x$. Это выражение можно рассматривать как искомую квазианалитическую зависимость $x_n = F(x_1, ..., x_{n-1})$.

Предлагаемый алгоритм

Из описания базового подхода следуют первоочередные направления его развития.

Их два:

— более точное приближение набора точек в пространстве главных компонент;
— автоматизация процесса аппроксимации.

Увеличение точности приближения разумно достичь, используя вместо плоскостей поверхности второго порядка. Так, представляется разумным использование эллиптических параболоидов в качестве приближающих поверхностей для данных, полученных в [1] после перехода в пространство главных компонент.

Автоматизация процесса подгонки семейства поверхностей к исследуемому набору точек становится тем более актуальной при использовании поверхностей второго порядка в качестве приближающих поверхностей, так как визуально-ручной подбор подходящей квадратичной поверхности представляется затруднительным.

Таким образом можно сформулировать следующую задачу для развития исходного подхода к анализу многомерных данных: необходимо по набору точек в трёхмерном пространстве, принадлежащих одной или нескольким поверхностям, определить явные уравнения этих поверхностей. При этом полезным может быть полуавтоматический режим, когда пользователь выбирает вид приближающей поверхности; в этом случае следует произвести наилучшую подгонку точек поверхностями выбранного типа.

Поставленную задачу разумно разбить на два этапа:

1. сегментация набора точек на множества точек, принадлежащих различным поверхностям;
2. определение уравнения для каждой из поверхностей.

Эти этапы будут подробно рассмотрены далее.

Последовательность шагов предлагаемого алгоритма представлена на рис. 1.

Сегментация набора точек.

Сегментация набора точек в трёхмерном пространстве является первым этапом обработки полученных после понижения размерности данных. Входными данными для алгоритма сегментации является набор точек в трёхмерном пространстве. О наборе точек можно сделать следующие предположения:

— точки образуют набор поверхностей, находящихся на некотором расстоянии друг от друга;

— поверхности не пересекаются.

Рассматриваемый в [1] класс задач предполагает, что поверхности оказываются похожими на эллиптический параболоид; кроме того поверхности имеют небольшую выпуклость. Пример входных данных показан на рис. 2.

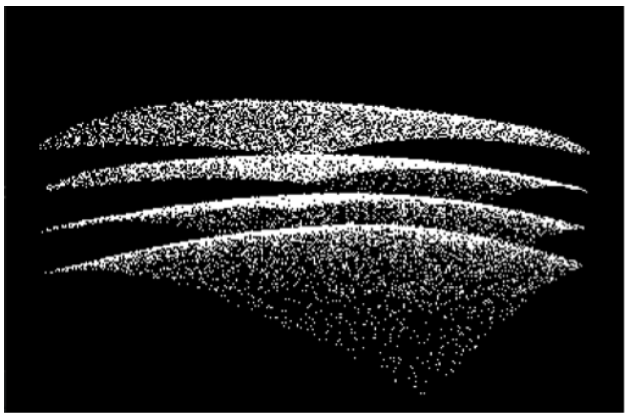

Рис. 2: Пример входных данных для алгоритма сегментации.

В случае соответствия входных данных описанными предположениям целью этапа сегментации является разделение набора точек на подмножества (кластеры), включающие в себя точки только одной поверхности. При этом оценка качества сегментации и возможности дальнейшего использования полученных результатов производится визуально. Предполагается, что несоответствие входных данных описанным предположениям устанавливается вручную; задача автоматического определения несоответствия в данной работе не рассматривалась.

Из описанных предположений о входных данных следует, что сегментацию произвести тем сложнее, чем ближе находятся поверхности друг к другу. Предельным является расстояние d_{min} между поверхностями: алгоритм выделяет правильно все поверхности, если расстояние между ними равно $d \geqslant d_{min}$ и не может определить поверхности, если расстояние между ними равно $d < d_{min}$. Тестирование алгоритма сегментации происходит на синтетических данных, состоящих из нескольких (например, 4) эллиптических параболоидов, заданных одной формулой и смещенных относительно друг друга на расстояние d по оси OZ (для задания параболоидов используется формула $z = \frac{x^2}{a^2} + \frac{y^2}{b^2}$).

Разумной метрикой для сегментации набора точек на множество поверхностей является гладкость локальной окрестности каждой из точек исходных данных. При построении алгоритма сегментирования необходимо найти компромисс между излишним и недостаточным сегментированием, выражающихся в слишком большом и слишком маленьком количестве сегментов соответственно. Излишнее сегментирование характерно для алгоритмов, основанных на анализе кривизны поверхности и производных высокого порядка.

В данной работе для кластеризации набора точек был применён модифицированный алгоритм выделения областей путём их наращивания [5]. Этот алгоритм основан на анализе значений кривизны и векторов нормалей к точкам рассматриваемого набора данных.

Вместе с набором точек, значениями локальной кривизны в точках набора и векторами нормалей в этих точках алгоритм принимает на вход два пороговых значения: максимальная разница углов нормалей и максимальное разница значений локальной кривизны. Точки рассматриваемого набора обрабатываются в определённом порядке, связанном со значениями локальной кривизны в точках набора. После добавления новой точки в текущий кластер рассматриваются «соседи» этой точки; соседние точки, удовлетворяющие обоим пороговым значениям, также добавляются в кластер.

Пример результата работы алгоритма сегментации показан на рис. 3. Алгоритм успешно разделил поверхности на 4 набора. Незначительное количество точек (помечены красным цветом) не были отнесены ни к одному из кластеров.

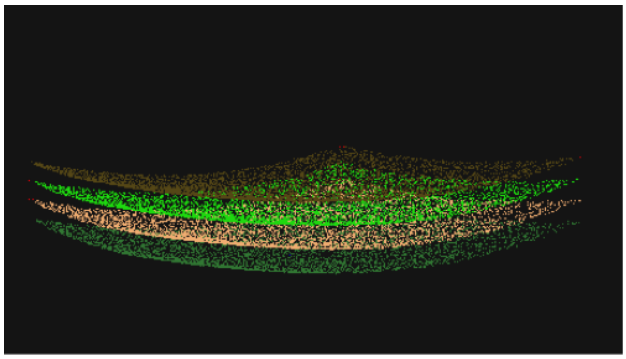

Рис. 3: Визуализация результата применения алгоритма сегментирования к набору точек, принадлежащих четырём параболоидам.

Апробация описанного алгоритма позволила сделать вывод об ограничениях, которым должны удовлетворять входные данные. Как было описано выше, интерес представляет минимальное расстояние d_{min} между подобными поверхностями, при котором алгоритм сегментирования корректно разделяет данные на кластеры. В результате экспериментов было установлено, что для данных, похожих на показанный на рис. 2 пример (несколько эллиптических параболоидов на некотором расстоянии друг от друга), расстояние между поверхностями связано с расстоянием между точками одной

поверхности. Было получено следующее соотношение: $d_{min} = 2.6r$, где r — расстояние между точками поверхностей, которые нужно отделить друг от друга. Алгоритм сегментации успешно выделяет поверхности при $d \geqslant d_{min}$.

Кроме нахождения расстояния d_{min} была проверена устойчивость процедуры сегментации к зашумленности входных данных. Было установлено, что максимальная амплитуда шума вдоль оси OZ, при котором возможно успешное разделение набора точек на поверхности, составляет $A_{max} = 0.35d$.

Нахождение уравнения поверхности второго порядка.

После того, как исходное множество точек было разбито на подмножества точек, принадлежащих разным поверхностям, необходимо найти явные выражения для каждой из поверхностей. Фактически, необходимо найти коэффициенты $c_0, ..., c_9$ в общей записи уравнения поверхности второго порядка: $f(c,p) = c_0 + c_1p_x + c_2p_y + c_3p_z + c_4p_x^2 + c_5p_y^2 + + c_6p_z^2 + c_7p_xp_y + c_8p_xp_z + c_9p_yp_z = 0$

Важно, что множество точек, составляющих каждую из поверхностей, могут не принадлежать одной поверхности второго порядка; при этом необходимым является нахождение поверхности, описывающей набор точек наилучшим образом. Найденная поверхность не должна обязательно содержать каждую из точек, но хорошо описывать набор точек в целом.

При построении и тестировании алгоритма определения уравнения поверхности второго порядка к исходным данным можно применить шум, сместив точки относительно исходной поверхности. Ответ при этом не должен значительно отличаться от случая, определения уравнения поверхности без шума.

Необходимо также обратить внимание на то, что рассматриваемые наборы точек составляют лишь сегмент поверхности. Как показано в [6], один и тот же сегмент поверхности может быть корректно отнесён к разным поверхностям второго порядка. Поэтому может быть полезной возможность явного выбора вида поверхности, которой будет приближаться рассматриваемый набор точек.

Алгебраические методы подгонки часто используются для подгонки поверхностей второго порядка [4]. Общая идея заключается в приближении нелинейной функции ошибки аппроксимации выражением вида: $\frac{\sum_{i=0}^{n} ||f(c,p_i)^2||}{q(c)}$, где $q(c)$ — нормализующая функция. В данной работе была выбрана функция $q(c) = \sum_{i=0}^{n} ||\nabla_p f(c,p_i)||^2$, предложенная в методе Таубина [7]. Метод Таубина является эффективным алгоритмом подгонки квадратичных поверхностей [4]; кроме того, в [6] предложен способ определять с его помощью уравнения поверхностей второго порядка определённого вида, что может быть полезно в рассматриваемой задаче.

В ходе исследования была проверена устойчивость выбранного метода определения коэффициентов $c_0, ..., c_9$ к шуму, который был добавлен к искусственно сгенерированным тестовым данным. Допустимая амплитуда шума зависит от площади сегмента, для которого происходит аппроксимация и приблизительно составляет $A = 0.03S$, где S — площадь рассматриваемого сегмента.

Заключение

В статье было рассмотрено развитие подхода к обработке многомерных данных, предложенного в [1]. Было предложено увеличить точность анализа входных многомерных данных путём использования квадратичных поверхностей при анализе главных компонент рассматриваемых многомерных данных и автоматизировать процесс обработки. Были рассмотрены и опробованы алгоритмы, позволяющие реализовать предложенные идеи.

В ближайшее время планируется провести эксперименты с целью проверить практическую применимость предлагаемого подхода и рассмотренных алгоритмов для решения многомерных задач вычислительной газовой динамики и, возможно, других практических задач.

Литература

[1] *Бондарев А.Е.* Анализ многомерных данных в задачах вычислительной газовой динамики // Научная визуализация. – 2014. – Т.6, № 5, – С.59-66.

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

[3] *Зиновьев А.Ю.* Визуализация многомерных данных – Красноярск: изд. КГТУ, 2000. – 180 с.

[4] *Chernov N., Ma H.* Least squares fitting of quadratic curves and surfaces // Computer Vision, Editor S. R. Yoshida, Nova Science Publishers, 2011. – pp.285-302.

[5] *T. Rabbani, F. van den Heuvel, G. Vosselman* Segmentation of point clouds using smoothness constraint // Proceedings of the ISPRS Commission V Symposium Image Engineering and Vision Metrology, 2006. – Vol.XXXVI Part 5. – pp.248-253.

[6] *Andrews J., Séquin C.* Type-Constrained Direct Fitting of Quadric Surfaces // Computer-Aided Design and Applications, September 2013. – Vol.11, Issue 1 – pp.107-119.

[7] *Taubin, G.* Estimation Of Planar Curves, Surfaces And Nonplanar Space Curves Defined By Implicit Equations, With Applications To Edge And Range Image Segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 13, 1991. – pp.1115-1138.