

Лядов Максим Алексеевич

E-mail: lyadov2@rambler.ru.

392000, г. Тамбов, ул. Мичуринская, 112 Д, кв. 206.

Тел. 89107529594.

Фареа Г. Султан

E-mail: soltantver@yahoo.com.

392000, г. Тамбов, ул. Никифоровская, 36, общежитие № 2, ком. 213.

Тел.: 89050474067.

Frolov Sergej Vladimirovich

State-Owned Educational Establishment of Higher Vocational Education "Tambov State Technical University".

E-mail: sergej.frolov@gmail.com.

8, 103, Bazarnaya street, Tambov, 392000, Russia.

Phone: +79204817586.

Lyadov Maxim Alexeevich

E-mail: lyadov2@rambler.ru.

206, 112D, Michurinskaya street, Tambov, 392000, Russia.

Phone: +79107529594.

Farea G. Sultan

E-mail: soltantver@yahoo.com.

Room 213, dormitory 2, 36, Nikiforovskaya street, Tambov, 392000, Russia.

Phone: +79050474067.

УДК 004.92:004.94

К.А. Шаропин, О.Г. Берестнева, В.А. Воловоденко, О.В. Марухина
ВИЗУАЛИЗАЦИЯ МЕДИЦИНСКИХ ДАННЫХ НА БАЗЕ ПАКЕТА
NOVOSPARK*

Мощным средством анализа информации являются интерактивные средства модификации графических представлений. В работе представлены различные подходы к визуализации результатов экспериментальных исследований. Приведены примеры решения прикладных задач с использованием NovoSpark Visualizer.

Методы визуализации; многомерные данные; представление информации в виде графических образов.

К.А. Sharopin, O.G. Berestneva, V.A. Volovodenko, O.V. Maruhina
VISUALIZATION OF THE MEDICAL DATA ON THE BASIS OF PACKAGE
NOVOSPARK

A powerful tool for analyzing information are interactive tools update graphic representations. The paper presents different approaches to visualize the results of experimental studies. Examples of solutions of applied problems using NovoSpark Visualizer.

Methods of visualization; the multivariate data; representation of the information as graphic images.

* Работа частично поддержана грантом РФФИ (проект № 08-06-00313а).

Введение. Визуализация данных – задача, с которой сталкивается в своей работе любой исследователь. К задаче визуализации данных сводится проблема представления в наглядной форме данных эксперимента или результатов теоретического исследования. Традиционные инструменты в этой области – графики и диаграммы – плохо справляются с задачей визуализации, когда возникает необходимость изобразить более трех взаимосвязанных величин. Визуализация должна быть одной из процедур-предикторов, дополняющих и взаимно верифицирующих численно-аналитические процедуры принятия решений. Возможности инструментария визуализации (ИВ) определяются областями его применения и конкретизируются совокупностями управленческих задач, решаемых с его помощью. Особенности ИВ определяются уникальностью и актуальностью концептуальной модели, на которой базируется методология и методика его построения. Возможности и особенности ИВ определяют архитектуру его программно-алгоритмического комплекса и структуру подсистемы его информационного обеспечения.

Современные компьютерные технологии используют широкий спектр методов визуализации информации. Легкость построения графиков и диаграмм с помощью ЭВМ все заметнее меняет когнитивные навыки исследователя. Современные пакеты анализа информации позволяют строить сотни типов различных графиков и диаграмм. Исследователь может одним взглядом обнаружить особенности, выявить закономерности и аномалии в больших объемах информации.

Мощным средством анализа информации являются интерактивные средства модификации графических представлений. Особенно широко графические методы используются в разведочном анализе данных, позволяя выявлять закономерности в многомерных массивах информации.

Визуализация экспериментальных данных, представленных в виде числовых таблиц. В экспериментальных исследованиях описание многомерных объектов обычно представлено в виде числовых таблиц. Методы визуализации такого рода информации основаны, как правило, на переходе от многомерной к двумерной системе координат (рис. 1) (метод главных компонент [1], методы структурного упорядочения [2] и др.).

Рассмотрим алгоритм формирования координат объектов в методе начального упорядочения [2].

Для оценки рассогласования структур в R^L и R^2 вычисляется матрица $D_N(X) = [d_{nk}]_{1,1}^{N,N}$ взаимных расстояний d_{nk} между элементами X_n и X_k из выборки X :

$$d_{nk} = d(Y_n, Y_k) = \left[\sum_{l=1}^L (x_{nl} - x_{kl})^2 \right]^{1/2}.$$

В n -й строке такой матрицы записаны расстояния от некоторого n -го элемента X_n до всех остальных $(N-1)$ элементов множества $\{X_n\}_1^N$, а k -й столбец матрицы образован расстояниями от всех элементов множества $\{X_n\}_1^N$ до некоторого k -го элемента. Любую n -ю строку матрицы $D_N(X)$ можно рассматривать как результат упорядочения элементов $\{X_n\}_1^N$ относительно n -го элемента X_n путем отображения этого множества на числовую ось действительных чисел R_n^+ . Задавая на оси R_n^+ положение n -го элемента и принимая его за начало отсчета (точку Y_n , координата которой на оси R_n^+ равна нулю), можно упорядочить образы

$\{Y_n\}_{j_1}^N$ выборки X на оси R_n^+ относительно n -го элемента, используя в качестве меры упорядочения расстояние от элемента X_n до всех остальных $(N-1)$ элементов. Из точки $Y_n \in R_n^+$ (начало отсчета в R_n^+) построим перпендикулярно к оси R_n^+ другую числовую ось R_k^+ , при этом k -й элемент выборки X расположим в точке пересечения осей R_n^+ и R_k^+ , а на ось R_k^+ отобразим множество $\{X_n\}_{j_1}^N$, подобно тому как это было осуществлено для оси R_n^+ . Координаты элементов $\{Y_n\}_{j_1}^N$ на оси R_k^+ представляют собой расстояния от k -го элемента до всех остальных $(N-1)$ элементов и позволяют судить о группируемости векторов $\{X_n\}_{j_1}^N$ около вектора X_k . Эти две оси R_n^+ и R_k^+ определяют некоторую псевдоплоскость $(R^+)^2$.

Таким образом, выбирая две любые строки (или два столбца) матрицы $D_N(X)$, можно образовывать новые псевдопространства $(R^+)^2$ образов $\{Y_n\}_{j_1}^N$ множества $\{X_n\}_{j_1}^N$. Полученная проектированием множества $\{X_n\}_{j_1}^N$ в $(R^+)^2$ совокупность образов $\{Y_n\}_{j_1}^N$ используется в качестве начального приближения для итерационной процедуры.

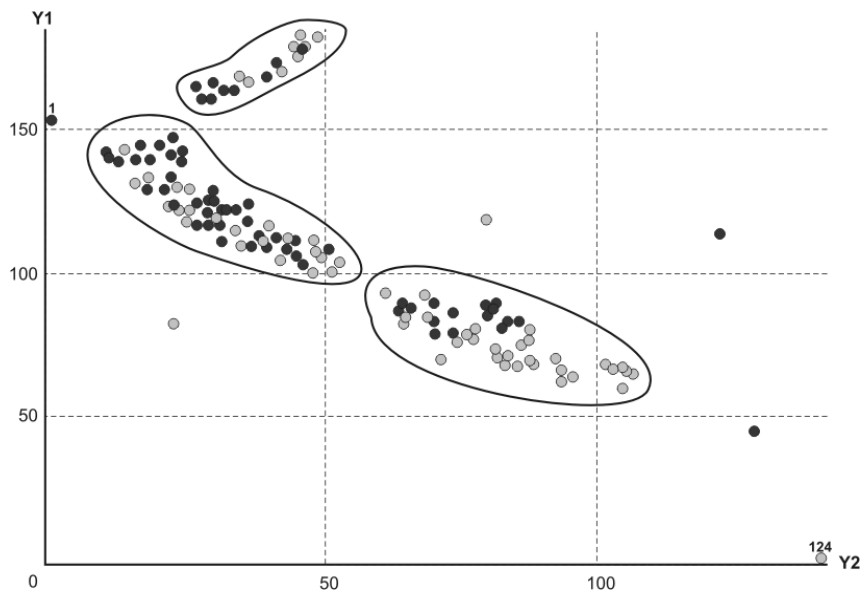


Рис. 1. Отображение многомерных экспериментальных данных (переменных психофизиологического состояния различных групп беременных женщин) в новой системе координат

Данный подход был использован авторами для визуализации экспериментальных данных в информационной системе оценки и мониторинга психофизиологического состояния беременных женщин [3].

Эффективность данного метода зависит от “хорошего” выбора строк матрицы $D_N(X)$, который не должен быть совсем случаен. Выбор близких в R_L элементов

X_n и X_k в качестве центров упорядочения остальных $(N-1)$ элементов на осях R_n^+ и R_k^+ нерационален, так как не дает существенно новой информации об упорядоченности выборки X , так что необходимо выбирать элементы X , относительно удаленные друг от друга. На рис. 1 представлен пример графического отображения многомерных данных, когда в качестве центров упорядочения нами выбраны «эталонный» объект (№1) (т.е. объект с наилучшими значениями переменных) и объект, имеющий «наихудшие» параметры (№124).

Визуализация многомерных экспериментальных данных на базе пакета *NovoSpark Visualizer*. Основной задачей визуализации данных является задача получения визуального образа, однозначно соответствующего набору данных. В пакете *NovoSpark Visualizer* (www.novospark.com) реализован предложенный В.А. Воловоденко подход, который позволяет провести визуализацию основных линейных конструкций: отрезок, ломаная линия, симплекс в многомерных пространствах. В рассматриваемом случае наиболее общей формой представления данных является вектор конечномерного пространства R_n :

$$A = (a_0, a_1, a_2, \dots, a_{n-1}) \in R_n. \quad (1)$$

Для перехода от данного вектора к визуальному образу будет использоваться базис ортонормированных функций $\{\varphi_i(\tau)\}_{i=0}^{\infty}$. В качестве такого базиса можно использовать известные функции, в частности ортонормированные полиномы Лежандра на отрезке $[0,1]$, множество которых мы обозначим через $\{l_i(\tau)\}_{i=0}^{\infty}$. В этом случае точке с координатами $A = (a_0, a_1, a_2, \dots, a_{n-1})$ можно поставить в соответствие функцию

$$F_A(\tau) = \sum_{i=0}^{n-1} a_i l_i(\tau). \quad (2)$$

Разница между формулами (1) и (2) заключается в том, что для вектора A из (1) возможно только аналитическое представление, в то время как для функции $F_A(\tau)$ возможно представление в виде графика этой функции. График функции – это структура, которая допускает визуальное представление на $[0,1]$. Между (1) и (2) устанавливается однозначная связь в обе стороны, т.е. взаимно-однозначная связь. Если ввести в рассмотрение второй вектор

$$B = (b_0, b_1, b_2, \dots, b_{n-1}) \in R_n, \quad (3)$$

то ему ставится в соответствие функция

$$F_B(\tau) = \sum_{i=0}^{n-1} b_i l_i(\tau). \quad (4)$$

Будем считать, что функции $F_A(\tau)$ и $F_B(\tau)$ являются визуальными образами точек A и B , принадлежащих пространству R_n (рис. 2).

Если с векторами A и B связать точки их концов, а начала векторов разместить в начале координат пространства R_n , то получается взаимно-однозначная связь между точками пространства R_n и функциями-образами $F_A(\tau)$ и $F_B(\tau)$ (рис. 3,б).

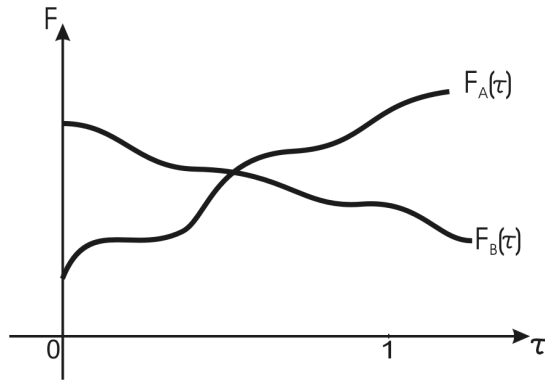


Рис. 2. Визуальные образы точек A и B, принадлежащих пространству R_n

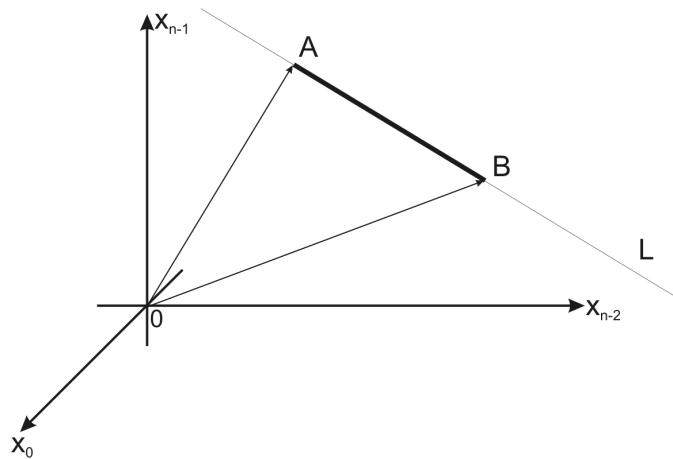


Рис. 3. Условное представление прямой L, проходящей через точки A и B в пространстве R_n

Введем переменную Z и образуем линейное выражение

$$C(z) = (1-z)A + zB = ((1-z)a_0 + zb_0, (1-z)a_1 + zb_1, \dots, (1-z)a_{n-1} + zb_{n-1}). \quad (5)$$

Очевидно:

$$C(0) = A \text{ и } C(1) = B.$$

Эта возможность дает право на рассмотрение функции $C(z)$ как функции, представляющей многомерную прямую линию, проходящую через точки A и B в пространстве R_n .

Следовательно, мы можем рассматривать выражение, аналогичное (5), для представления отрезка AB:

$$AB = (1-z)A + zB, \quad z \in [0,1]. \quad (6)$$

Из выражения (6) следует, что можно образовать функцию

$$F_{AB}(z) = \sum_{i=0}^{n-1} (1-z)a_{i_i}(\tau) + zb_{i_i}(\tau). \quad (7)$$

Обратим внимание, что эта функция будет зависеть от двух аргументов $\{z, \tau\}$. Это очень важно, так как дает возможность получения графика функции $F_{AB}(z) = F_{AB}(z, \tau)$, который является визуальным образом отрезка АВ (рис. 4). Таким образом, получается, что на квадрате $[0,1] \times [0,1]$ можно построить гладкую поверхность по формуле (7), которая будет однозначно соответствовать аналитическому выражению (6), представляющему многомерный отрезок АВ.

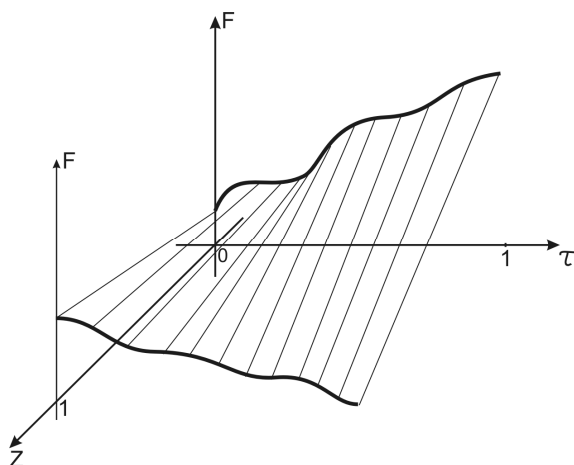


Рис. 4. Визуальное представление поверхности, соответствующей отрезку АВ

Таким образом, основой визуализационного подхода *NovoSpark* является представление *многомерного* наблюдения в виде *двумерного* образа – кривой. При этом гарантируется, что близким по значениям наблюдениям А и В будут соответствовать визуально близкие образы-кривые; для сильно различающихся по значениям наблюдений их образы-кривые будут заметно отличаться.

NovoSpark Visualizer (разработка компании *NovoSpark*) – это современный инструмент визуализации, который соединяет в себе интуитивность визуального анализа с описательной мощью статистических методов и предоставляет наиболее полную информацию о многомерных данных. Продукт позволяет легко манипулировать загруженными наборами данных и пересчитывать модель и ее графическое представление, давая возможность видеть результаты экспериментирования с данными. Используя этот инструмент, становится возможным автоматически классифицировать наблюдения, определять наиболее важные переменные в модели, производить кластеризацию данных, визуально сравнивать индивидуальные наблюдения и целые наборы данных, а также выполнять много других задач в работе с многомерными данными [4,5].

Заключение. Суть приведенных методов – повышение наглядности отображаемой медицинской информации. Результаты проведенных авторами экспериментальных педагогических и психофизиологических исследований позволяют уверенно утверждать, что визуализация является одним из наиболее перспективных направлений повышения эффективности методов анализа и представления многомерных медицинских данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Дюк В.А., Эммануэль В. Информационные технологии в медико-биологических исследованиях. – СПб.: Питер, 2003. – 528 с.

2. Попечителев Е.П., Старцева О.Н. Аналитические исследования в медицине, биологии и экологии: Учебное пособие. – М.: Высшая школа, 2003. – 279 с.
3. Берестнева О.Г., Добрянская Р.Г., Муратова Е.А., Шаропин К.А. Интеллектуальная система выявления групп риска среди беременных женщин // Информатика и системы управления. – 2008. – № 2 (16). – С. 22-23.
4. Ротарь В.Г., Лукьянец А.А., Татарников В.А., Эйдензон Д.В. Особенности визуализации многомерных данных средствами Novospark Visualizer для бенчмаркинга теплоснабжающих предприятий // Информационные и математические технологии в науке и управлении: Труды XIV Байкальской Всероссийской конференции. – Иркутск, 5-15 июля 2009. – Иркутск: ИСЭМ СО РАН, 2009. – С. 80-90.
5. Берестнева О.Г., Марухина О.В., Шаропин К.А., Хантарина С.А., Шевцова А.В. Визуализация результатов экспериментальных исследований в социальной сфере // Информационные системы и технологии (ИСТ-2010): Материалы XVI Международной научно-технической конференции. – Нижний Новгород, 23 апреля 2010. – Нижний Новгород: НГТУ, 2010.

Шаропин Константин Александрович

Национальный исследовательский Томский политехнический университет.

E-mail: kashar@mail.ru.

634004, г. Томск, ул. Советская, 84/3.

Тел.: 83822426100.

Берестнева Ольга Григорьевна

E-mail: ogb@rambler.ru.

Воловоденко Виталий Алексеевич

E-mail: vav@osu.cctpu.edu.ru.

Тел.: 83822420459.

Марухина Ольга Владимировна

E-mail: marukhina@tpu.ru.

Sharopin Konstantin Aleksandrovich

National Research Tomsk Polytechnic University.

E-mail: kashar@mail.ru.

84/3, Sovetskaya street, Tomsk, 634004, Russia.

Phone: +73822426100.

Berestneva Olga Grigorevna

E-mail: ogb@rambler.ru.

Volovodenko Vitaliy Alexeevich

E-mail: vav@osu.cctpu.edu.ru.

Phone: +73822420459.

Maruhina Olga Vladimirovna

E-mail: marukhina@tpu.ru.