мости от уровня использования ресурсов эффективно монетизировать предоставленные сервисы.

Заключение. Таким образом, в контексте организации единого центра обработки данных о дорожной ситуации необходимо рассматривать применение современных технологий обработки больших данных, а именно внедрение систем хранения данныхи единой систему управления и оркестрации, что позволит централизованно собирать и обрабатывать как структурированные, так и не структурированные данные. Все это даст возможность формирования хранилища данных для дальнейшего анализа и прогнозирования дорожных ситуаций и принципиального развития и реорганизации транспортной-дорожной системы города. Исследования проводятся в рамках проекта AP05133699 «Исследование и разработка инновационно-телекоммуникационных технологий с использованием современных кибертехнических средств для интеллектуальной транспортной системы города».

ЛИТЕРАТУРА

1. Миляр А. Умный и безопасный город. [Электронный ресурс]. URL: http://www.jetinfo.ru/stati/umnyj-i-bezopasnyj-gorod. (дата обращения: 21.05.2019).

2. Соммер А. Кибернетический оркестр. [Электронный ресурс]. URL: https://xakep.ru/2018/10/08/kubernetes-docker/ (дата обращения: 21.05.2019).

3. Docker: оркестрация. [Электронный ресурс]. URL: https://ast.rocks/blog/docker-orchestration (дата обращения: 21.05.2019).

4. Оркестровка (ИТ) [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/%D0%9E%D1%80%D0%BA%D0%B5%D1%81%D1%82%D1%80%D0%BE%D0%B2%D0%BA%D0%B0_(%D0%98%D0%A2) (дата обращения: 22.05.2019).

5. Это уже явно не фантастика. [Электронный ресурс]. URL: http://www.jetinfo.ru/stati/eto-uzhe-yavno-ne-fantastika (дата обращения: 22.05.2019).

6. Сервис-ориентированная архитектура. [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/Сервис-ориентированная_архитектура (дата обращения: 22.05.2019).

7. Что такое оркестрация контейнеров [Электронный ресурс]. URL: https://www.xelent.ru/blog/chto-takoe-orkestratsiya-konteynerov/ (дата обращения: 22.05.2019).

8. КИСУ ГППТ. [Электронный ресурс]. URL: http://orgp.spb.ru/kisu_gppt/ (дата обращения: 23.05.2019).

9. Интеллектуальные транспортные системы [Электронный ресурс]. URL: https://studref.com/361389/tehnika/intellektualnye_transportnye_sistemy (дата обращения: 23.05.2019).

10. Центры обработки данных Повышение масштабируемости, гибкости и безопасности бизнеса. [Электронный ресурс]. URL: https://becsys.ru/uploads/files/solutions/technological-solutions/3/Business_Ecosystems_Data_Centers.pdf (дата обращения: 24.05.2019).

ВИЗУАЛИЗАЦИЯ МНОГОМЕРНЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ КРИВЫХ ЭНДРЮСА

А. М. Ширыкалов
(г. Томск, Томский политехнический университет)
e-mail: ams28@tpu.ru

VISUALIZATION OF MULTIDIMENSIONAL DATA WITH ANDREWS CURVES

A. M. Shirykalov
(Tomsk, Tomsk Polytechnic University)

Abstract. In a recent decades processing power of computer systems has had a significant surge. So now we are capable of processing that huge amount of data which has been collected by various information

systems around the globe since first databases were implemented. A task of multidimensional data analysis and processing is an important area of data science, as no matter what your data describes, it is likely to have more than one parameter. Crucial part of any data analysis is its visualization. It helps researcher understand what kind of data he is working with, does it split into any classes, does it contain any outliers and so on. In this paper the use of Andrews curves in multidimensional data visualization as part of its primary analysis is described. As an example, visualization of Wheat Seeds Dataset via Andrews curves is given.

Keywords: Andrews curves, multidimensional data visualization, multidimensional data outliers, primary data analysis, Wheat Seeds Dataset.

Введение Визуализация является важной частью первичной обработки данных, так как позволяет исследователю произвести их качественный анализ. Существует множество различных методов визуализации, однако многие из них становятся малоэффективными при увеличении размерности данных. Так, например, чтобы визуализировать корреляцию между всеми парами параметров для данных с количеством параметров N, равным 25, необходимо построить

$$\frac{N(N-1)}{2} = 25 \times 12 = 300$$

диаграмм рассеяния. Одним из методов, позволяющих представить многомерные данные на плоскости, является построение кривых Эндрюса.

Кривые Эндрюса Главной особенностью метода визуализации многомерных данных, описанного Эндрюсом в его работе [1], является возможность представить данные любой размерности в виде кривых на плоскости. При этом каждой записи в данных ставится в соответствие функция в виде ряда Фурье:

$$f_x(t) = x_1 2^{-\frac{1}{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \dots,$$

где:
$x(x_1, x_2, \dots x_m)$ – запись (точка) из набора данных,
$x_i (i = 1, \dots, m)$ – изменяемые переменные,
m – размерность данных.

Полученные кривые строятся на плоскости в промежутке $-\pi < t < \pi$ и обладают рядом важных свойств. Во-первых, для любых точек x и y

$$f_x(t) + f_y(t) = f_{x+y}(t),$$

откуда следует, что множество полученных кривых обладает статистическими характеристиками, сходными с таковыми у множества данных. Среди них: среднее значение (1), расстояния между элементами (2).

$$f_{\bar{x}}(t) = \frac{1}{n}\sum_{i=1}^{n} f_{x_i}(t), \tag{1}$$

где $\bar{x}$ – средний вектор,

$$\left\| f_x(t) - f_y(t) \right\|_{L_2} = \int_{-\pi}^{\pi} \left[f_x(t) - f_y(t) \right]^2 dt = \pi \sum_{i=1}^{m} (x_i - y_i)^2 = \pi \|x - y\|^2. \tag{2}$$

Так же, при условии, что параметры $x_i (i = 1, \dots, m)$ являются некоррелируемыми случайными величинами с дисперсией σ^2, справедливо

$$D[f_x(t)] = \sigma^2 (2^{-1} + \sin^2(t) + \cos^2(t) + \dots) =$$

$$= \begin{cases} \frac{1}{2}\sigma^2 m, & \text{если } m \text{ нечетно,} \\ \frac{1}{2}\sigma^2 \left[m - 1 + 2sin^2\left(\frac{mt}{2}\right) \right], & \text{если } m \text{ четно.} \end{cases}$$

Очевидно, что независимо от значения m выполняется следующее неравенство

$\frac{1}{2}\sigma^2(m-1) < D[f_x(t)] < \frac{1}{2}\sigma^2(m+1)$.

Эти свойства позволяют, полученный график, делать следующие предположения относительно данных:

- Если несколько кривых находятся близко друг к другу при всех значениях t, то точки данных, соответствующие этим кривым, близки в соответствии с Евклидовой метрикой. Такая группа кривых отображает кластер точек данных. С другой стороны, если кривая визуально сильно отличается от большинства других, соответствующая точка, возможно, является выбросом.

- Если несколько кривых находятся близко друг к другу при определенных значениях t_i, то точки данных, соответствующие этим кривым, близки в направлениях, описываемых векторами

$\mathbf{f}_1(t_i) = (2^{-\frac{1}{2}}, \sin(t_i), \cos(t_i), \sin(2t_i), \dots)$.

Это может позволить определить кластеры данных даже с присутствием дополнительных параметров [1].

- Если точка данных $\boldsymbol{y}$ лежит на прямой, соединяющей точки $\boldsymbol{x}$ и $\boldsymbol{z}$, тогда для всех значений t, $f_y(t)$ находится между $f_x(t)$ и $f_z(t)$.

Таким образом, построение кривых Эндрюса позволяет ответить на следующие вопросы о данных:

- Содержат ли данные ярко выраженные классы и кластеры?
- Содержат ли данные явные выбросы?

Если данные уже разделены на классы:

- Чем классы схожи между собой, а чем отличаются?

Для более эффективного применения данного метода визуализации можно произвести предварительную подготовку данных, в которую входит их нормализация, выделение главных компонент, исключение сильно связанных параметров [1, 2].

Визуализация набора данных *Wheat Seeds Dataset* с помощью кривых Эндрюса

Набор данных *Wheat Seeds Dataset* [3] состоит из данных о 210 зернах пшеницы трех различных видов: *Kama*, *Rosa* и *Canadian*, по 70 штук каждого. Для каждого ядра определены следующие параметры: площадь на фотографии (*Area)*, периметр на фотографии (*Perimeter*), компактность (*Compactness*), вычисленная через площадь и периметр по формуле

$Compactness = \frac{4\pi Area}{Perimeter^2}$,

длина (*Kernel.*Length), ширина (*Kernel.*Width), коэффициент асимметрии (*Asymmetry.Coeff*), длина паза (*Kernel.Groove*) и вид пшеницы.

Так как параметр *Compactness* функционально связан с параметрами *Area* и *Perimeter*, при построении графиков он не учитывался. Также для демонстрации кластеризации кривых в приведенных ниже рисунках не учитывался вид пшеницы. Ниже, на рисунках 1 и 2 представлены кривые Эндрюса для ненормализованных и нормализованных предложенным в [2] способом данных соответственно.

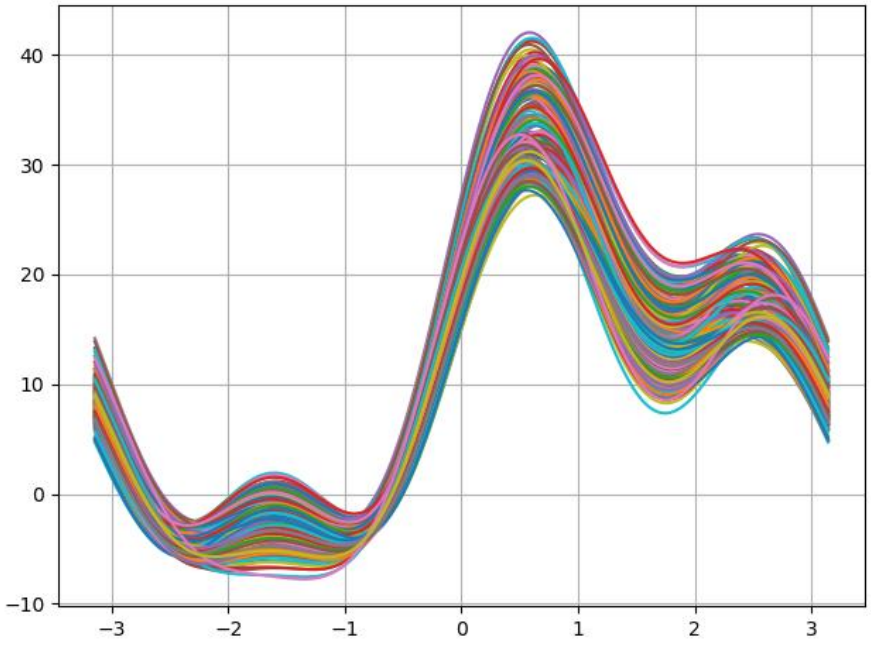

Рисунок 1. Кривые Эндрюса для ненормализованных данных

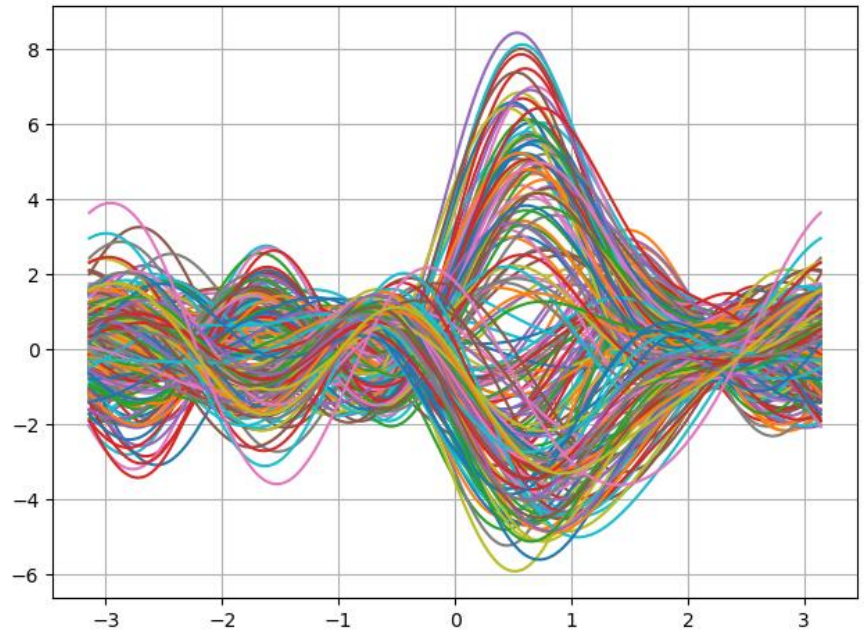

Рисунок 2. Кривые Эндрюса для нормализованных данных

На рисунке 2, в отличие от рисунка 1, хорошо видны три группы схожих кривых, соответствующих разным видам пшеницы. Это показывает необходимость нормализации данных перед применением данного метода визуализации. На графике отсутствуют кривые, сильно отдаленные ото всех групп, следовательно в данных отсутствуют явные выбросы.

Зная, к какому виду принадлежит каждое зерно, можно построить кривые Эндрюса для средних векторов каждого из классов (Рис. 3).

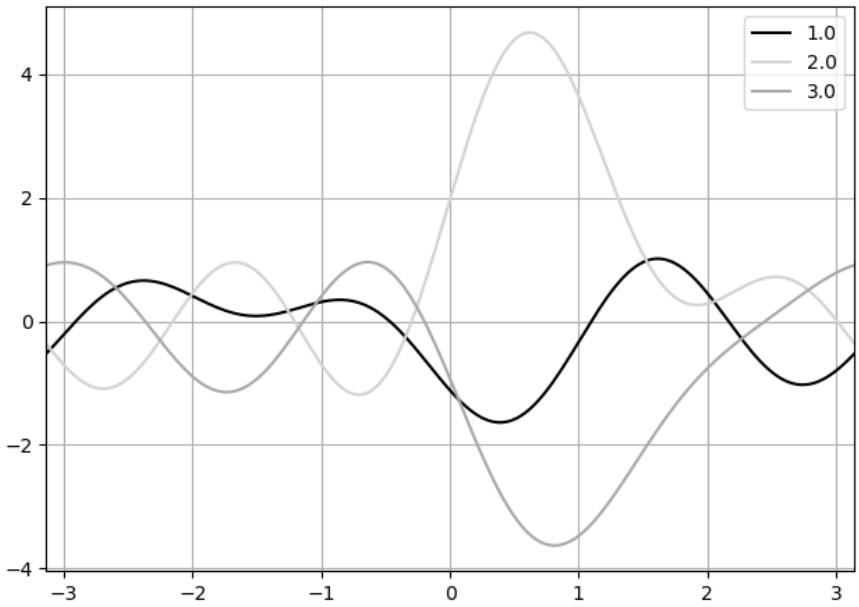

Рисунок 3. Кривые Эндрюса для средних векторов видов семян

На рисунке 3 номера 1, 2 и 3 соответствует видам *Kama*, *Rosa* и *Canadian* в таком порядке. Можно заметить, что кривые видов 2 и 3 расположены асимметрично относительно оси абсцисс. Учитывая, что данные были нормализованы с использованием формулы

$$x_k{}' = (\frac{x_{ki} - \overline{x_i}}{\sigma_i}(i = 1, \dots, m)),$$

можно сделать вывод о том, что в среднем, значения параметров у классов 2 и 3 расположены асимметрично относительно средних значений этих параметров.

Так же можно увидеть, что кривая класса 1 имеет меньшую амплитуду, чем кривые классов 2 и 3, а следовательно, значения параметров этой группы расположены ближе к средним значениям этих параметров.

ЗаключениеВ работе было показано применение кривых Эндрюса для визуализации многомерных данных, описаны их основные свойства. С помощью данного метода был визуализирован набор данных *Wheat Seeds Dataset*, на основе визуализации проведен анализ структуры данных, сделаны выводы о наличие в нем выбросов, классов, отношении между классами.

ЛИТЕРАТУРА

1. Andrews, D. F. Plots of High-Dimensional Data // Biometrics. – 1972. – Т. 28. – № 1 – С. 125-136.
2. Грошев С.В., Пивоварова Н.В. Использование кривых Эндрюса для визуализации многомерных данных в задачах многокритериальной оптимизации // Машиностроение и компьютерные технологии. – 2015. – № 12 – С. 197-214.
3. M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, S. Zak A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images // Information Technologies in Biomedicine. – 2010. – С. 15-24.