ТЕХНИЧЕСКИЕ НАУКИ — МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ЧИСЛЕННЫЕ МЕТОДЫ И КОМПЛЕКСЫ ПРОГРАММ (05.13.18)

05.13.18

А.А. Голубничий, А.Д. Яблонцева

Хакасский государственный университет имени Н.Ф. Катанова,
инженерно-технологический институт,
кафедра программного обеспечения вычислительной техники и автоматизированных систем,
Абакан, artem@golubnichij.ru

РАЗРАБОТКА СИСТЕМЫ АВТОМАТИЗАЦИИ РАЗВЕДОЧНОГО АНАЛИЗА ДАННЫХ

В работе описывается процесс разработки программного продукта, позволяющего автоматизировать разведочный анализ данных посредством многопанельной графики, реализованной в пакете brinton [1]. Анализируются варианты реализации разведочного анализа при помощи языка программирования R. Описывается набор функций, реализованных в программном обеспечении. Приводится рендеринг отчета, генерируемого как результат работы приложения.

Ключевые слова: *разведочный анализ данных, многопанельная графика, прикладное программное обеспечение.*

Введение

Область разведочного анализа данных (EDA) существенно отличается от любой другой в классической системе анализа данных. Основной причиной этому служит факт того, что по завершении процедуры EDA не проверяется та или иная гипотеза, наоборот, предварительное проведение разведочного анализа данных помогает выдвинуть гипотезу в дальнейшем, в зависимости от того, что увидит аналитик. По этой причине область автоматизации разведочного анализа данных невозможна в части решения конкретной проблемы, однако автоматизация самого разведочного анализа данных достаточно тривиальна, что позволяет создать автоматизированную информационную систему.

Инструменты для автоматического создания графики и статистик для набора данных, называются автоматизированным разведочным анализом данных или autoEDA. Эти инструменты облегчают некоторые характерные задачи EDA, такие как описание переменных и подтверждение наблюдений или отношения, установленные между значениями одной или нескольких переменных.

Для разработки системы автоматизации проведения разведочного анализа данных, средствами языка программирования R, планируется использовать пакет brinton. Он показывает только графику, что позволяет его классифицировать как инструмент для автоматизированных графических исследовательских анализов данных или autoGEDA.

AutoGEDA и многопанельная графика

Важной особенностью пакета brinton является то, что он сочетает в себе различные графические типы, позволяющие строить множество графических элементов относящихся к одному набору данных в виде многопанельной графики [2]. Существует ряд инструментов autoGEDA, входящих в базовый функционал языка R и в пакеты, размещенные в официальном репозитории CRAN.

Среди пакетов R, реализующих автоматизацию разведочного анализа данных, лишь несколько позволяют строить изображения. К таким пакетам можно отнести tabplot, visdat и inspectdf. В частности, tabplot и visdat, по сути, предлагают варианты табличных диаграмм, в то время как inspectdf дает возможность строить графики в виде гистограмм.

Другой набор пакетов группирует переменные набора данных по типу и представляет распределение каждой переменной в ячейки многопанельного рисунка. К таким пакетам стоит отнести xray [3], DataExplorer [4] и SmartEDA [5]. Пакеты dataMaid и summarytools предлагают другой способ «наблюдать» за всеми переменными. Эти пакеты имеют функции, которые производят описательную сводку переменных вместе с гистограммой.

Пакеты AutoEDA обычно имеют двойное представление результатов: табличное и графическое. Некоторые из них, такие как dataMaid, summarytools и SmartEDA, позволяют создавать автоматические отчеты и даже адаптировать эти отчеты к потребностям конкретного пользователя.

Пакет brinton

Пакет brinton был создан для облегчения разведочного анализа данных. Основная идея состоит в том, чтобы помочь пользователю в использовании графиков с помощью трех функций: wideplot(), longplot() и plotup().

Функция wideplot() позволяет пользователю исследовать набор данных в целом, используя сетку для построения изображений в которой каждая переменная представлена в виде нескольких графиков. Сначала функция wideplot() группирует переменные в следующей последовательности: логическая, упорядоченная факторная, факторная, символьная, datetime и числовая. Затем она создает многопанельную графику в формате html, в которой каждый переменная набора данных представлена в строке сетки, а в каждом столбце отображается различная доступная графика для каждой переменной.

Единственный аргумент, необходимый для получения результата – это данные, в качестве возможных классов выступают data.frame, параметр dataclass выбирает и сортирует типы отображаемых переменных; ncol фильтрует первые n столбцы сетки, от 3 до 7, которые будут отображаться в итоговом отчете. Пример вызова функции wideplot() приведен на рисунке 1.

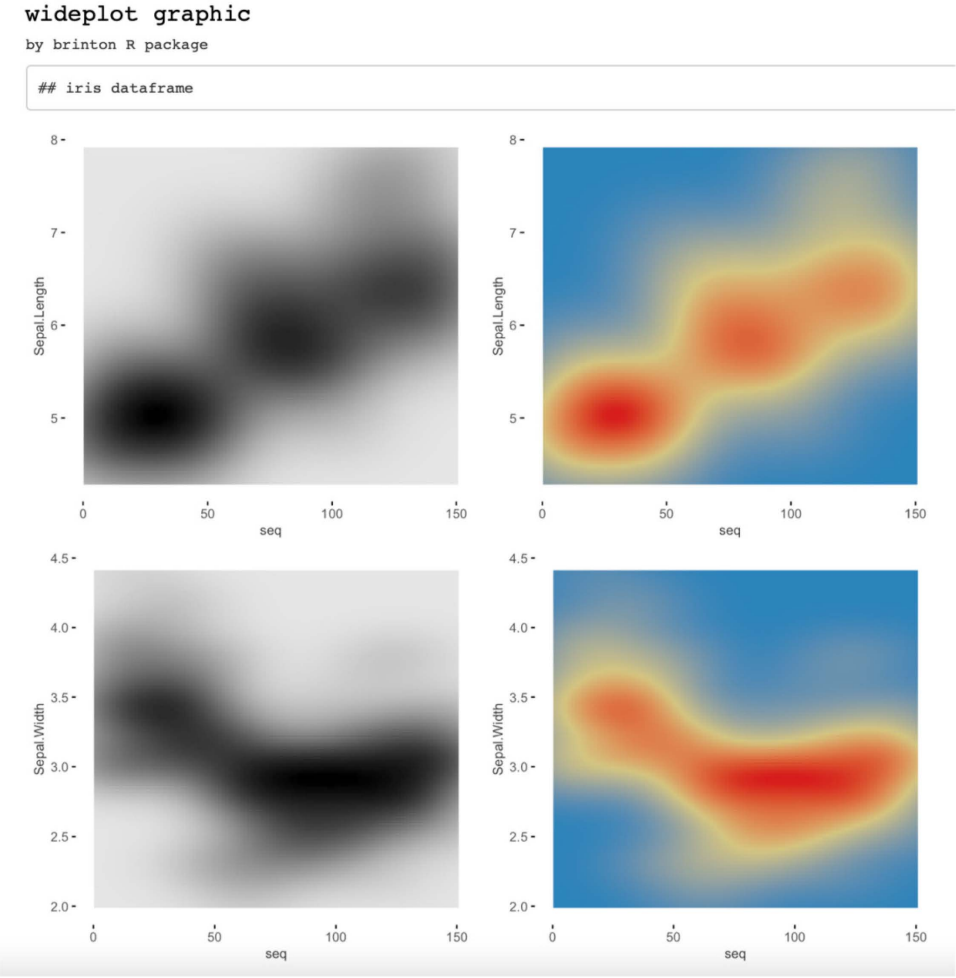

Рис. 1 – Пример реализации функции wideplot() на наборе данных Iris

После изучения набора данных в целом, можно использовать функцию longplot(), позволяющую исследовать другие графики для данной переменной. Эта функция также представляет собой сетку изображений, но вместо того, чтобы показывать выборку графиков для каждой переменной, она представляет полный спектр графиков, доступный в пакете, для представления одной переменной.

Графический тип называется longplot, потому что он показывает полный диапазон доступных графиков для представления отношений между значениями ограниченного набора переменных. В отличие от сетки функции wideplot, сетка функции longplot не включает параметры, ограничивающие диапазон отображаемой графики. Пример вызова функции longplot() приведен на рисунке 2.

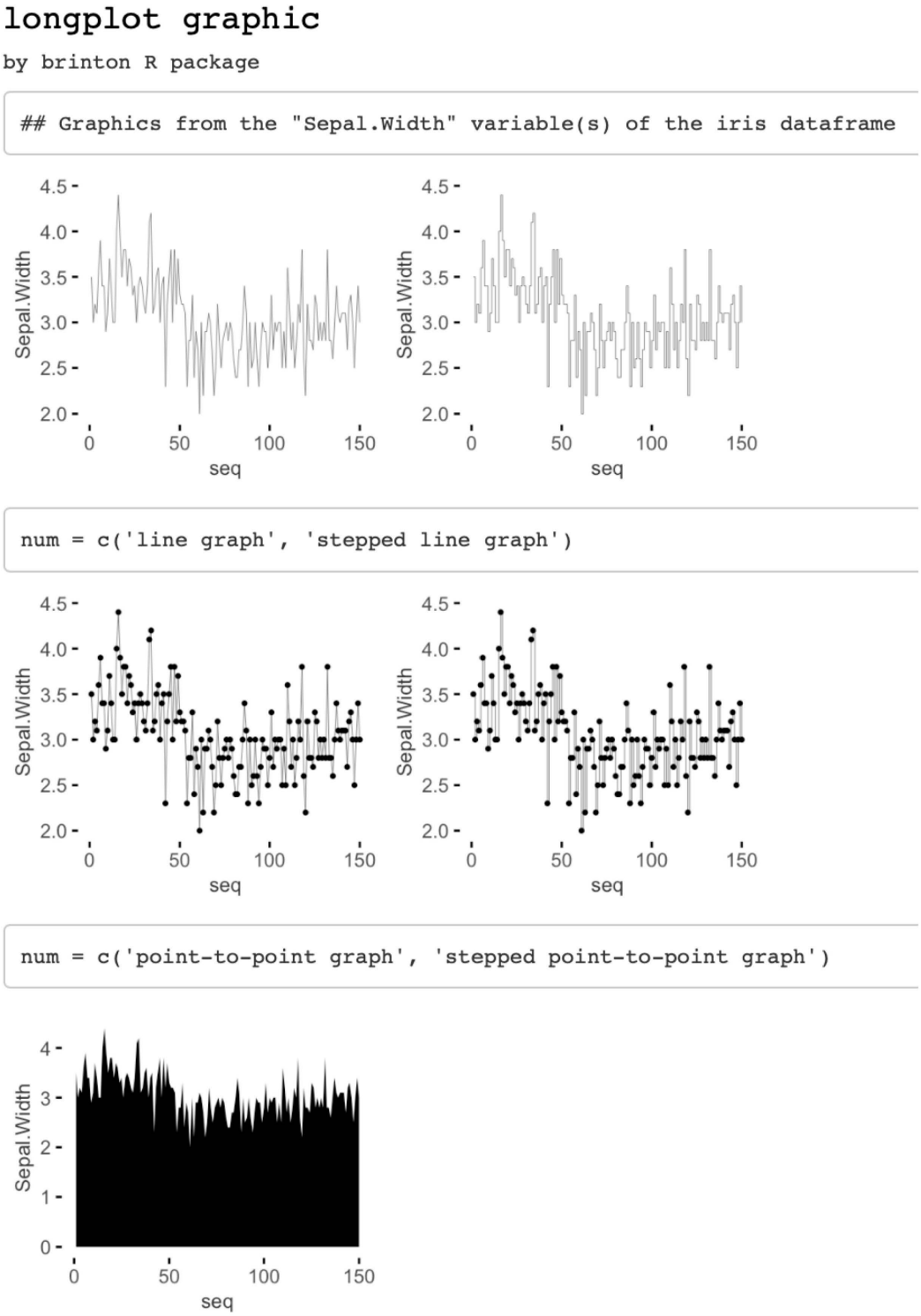

Рис. 2 – Пример реализации функции longplot() на наборе данных Iris

Пакет brinton основан в первую очередь на грамматике графики, реализованной в R пакетом ggplot2. Кроме того, он использует пакет gridExtra для создания многопанельной графика и rmarkdown, для динамического составления результатов.

Создание автоматизированного приложения

Для автоматизации работы с функциями пакета brinton и динамической генерации необходимых отчетов, было реализовано приложение, позволяющее быстро и без лишних дополнительных действий обрабатывать исходные данные вне зависимости от формата и строить соответствующие динамические отчеты в нужном формате.

Приложением позволяет загружать набор данных и в зависимости от выбранной функции выдавать необходимый файл отчета. Рендеринг приложения приведен на рисунке 3.

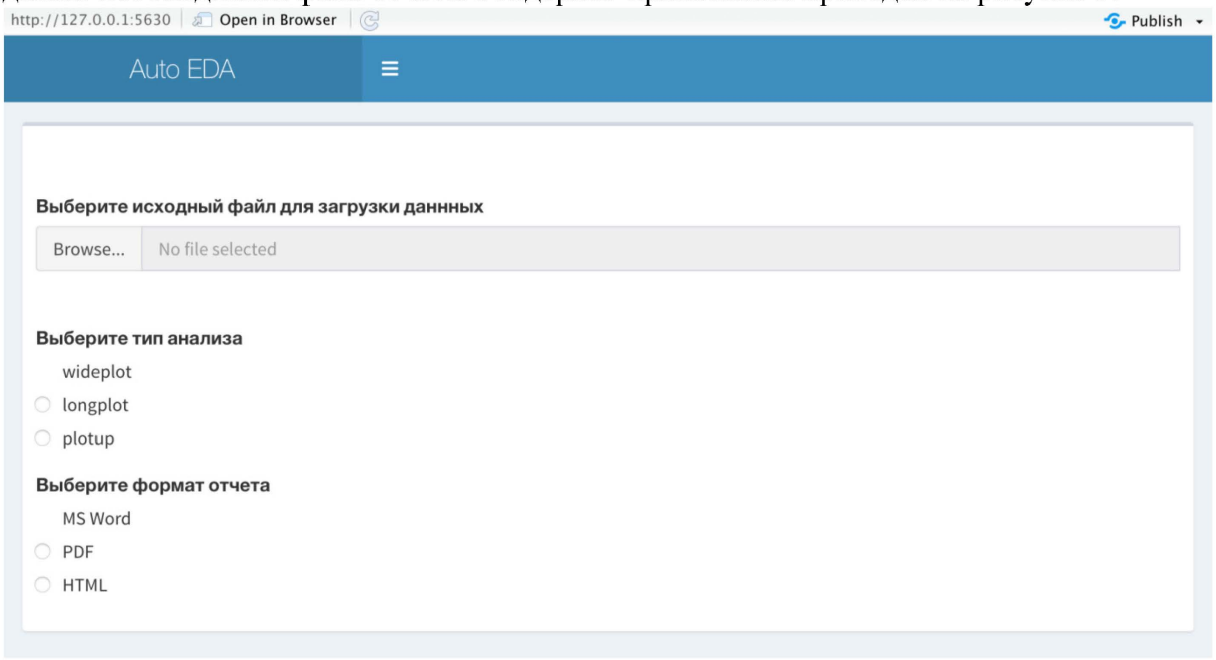

Рис. 3 – Рендеринг итогового приложения

Выводы

В рамках реализации исследования был проанализирован функционал пакета brinton, как инструмент графического представления EDA, предназначенный для облегчения представления, выбора и редактирования статистических графиков, построенных на ggplot2. Для упрощения процедура проведения разведочного анализа данных был разработан программный продукт, позволяющих создавать необходимые отчеты для последующего анализа и генерации гипотез на его основе.

Список литературы

1. CRAN – Package brinton [Электронный ресурс] URL: https://CRAN.R-project.org/package=brinton (дата обращения: 10.08.2021).
2. Data Visualization in R [Электронный ресурс] URL: https://www.datavis.ca/courses/RGraphics (дата обращения: 08.08.2021).
3. CRAN – Package xray [Электронный ресурс] URL: https://CRAN.R-project.org/package=xray (дата обращения: 10.08.2021).
4. CRAN – Package DataExplorer [Электронный ресурс] URL: https://CRAN.R-project.org/package=DataExplorer (дата обращения: 10.08.2021).
5. CRAN – Package SmartEDA [Электронный ресурс] URL: https://CRAN.R-project.org/package=SmartEDA (дата обращения: 10.08.2021).