Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

ggESDA: An R Package for Exploratory Symbolic Data Analysis using ggplot2

Bo-Syue Jiang
National Taipei University

Han-Ming Wu
National Chengchi University

Abstract

This paper presents the **ggESDA** package, which we developed for exploratory symbolic data analysis in R. Based on **ggplot2** Wickham (2009), the **ggESDA** package which is familiar programming structure with its parent provides a wide variety of graphical techniques such as histogram, 3D-scatterplot and radar plot. In addition, a general and customized transformation function `classic2sym()` is implemented for generating a symbolic data table from classical data frame by clustering algorithm, **RSDA** Rojas (2015) function and user-defined method. wait for edit......

Keywords: data visualization, symbolic data analysis, exploratory data analysis, **ggplot2** extensions, interval-valued data, R.

1. Introduction

"In Data Science the aim is to extract new knowledge from Standard, Big, and complex data. Often these data are unstructured with variables defined on different kinds of units. They can also be multi-sources (as mixtures of numerical and textual data, with images and networks)." Diday and Edwin (2018). The statement indicates that not only conventional data but the unstructured data, also known as symbolic data, is vital for data science. Rather than the classical data represented by a single value, symbolic data with measurements on p random variables can be p-dimensional statistical units such as hypercubes or histograms. The field of symbolic data analysis (SDA) Billard and Diday (2007) is to broaden the application aspects of statistical methodologies, extend traditional cognition of a form of data unit and build a brand-new analysis system of data science. Recent developments in the field of big data analytics have led to a renewed interest in complex structure data such as symbolic data. As shown in Figure 1, the number of researches in SDA represents an increasing trend from 1998 to 2020, which outstands the importance of it during the years. Among ScienceDirect,

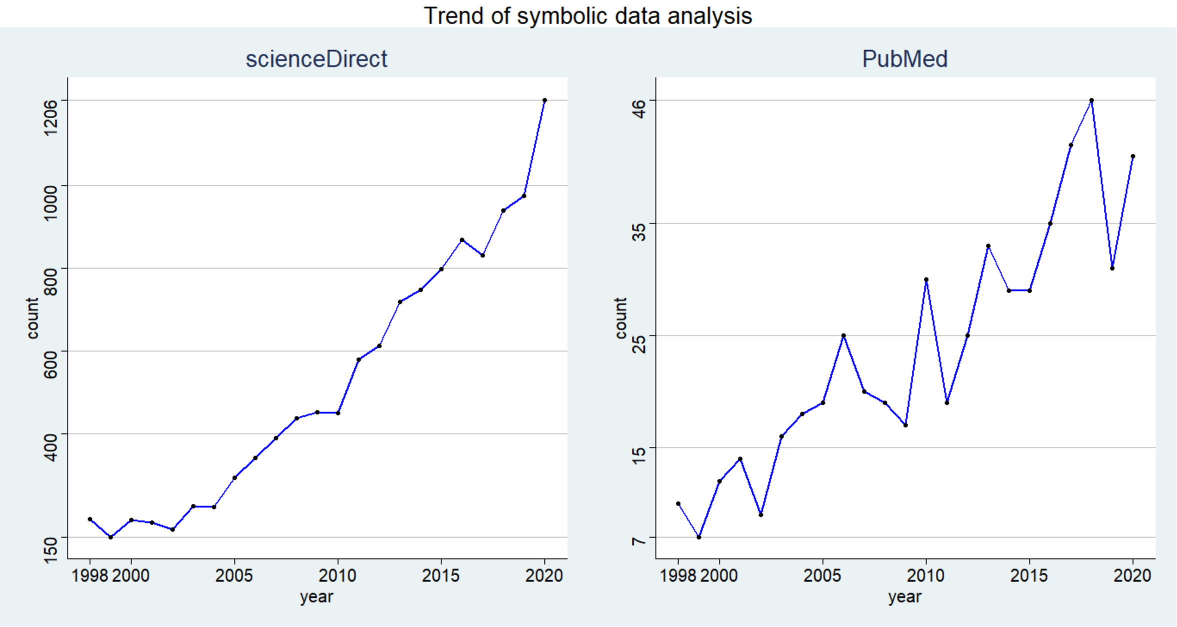

Figure 1: The number of "symbolic data analysis" or "interval-valued data" related articles in researches and applications according to PubMed and ScienceDirect online database over time from 1998 to 2020.

Engineering and Computer Science lead the subject areas obviously, shown in Figure 2.

In practice, the symbolic data is often generated by aggregating massive datasets into intervals in order to make the management easier. An interval-valued symbolic random variable

In practice, the symbolic data is often generated by aggregating massive datasets into intervals in order to make the management easier. An interval-valued symbolic random variableIn practice, the symbolic data is often generated by aggregating massive datasets into intervals in order to make the management easier. An interval-valued symbolic random variableIn practice, the symbolic data is often generated by aggregating massive datasets into intervals in order to make the management easier. An interval-valued symbolic random variableIn practice, the symbolic data is often generated by aggregating massive datasets into intervals in order to make the management easier. An interval-valued symbolic random variableIn practice, the symbolic data is often generated by aggregating massive datasets into intervals in order to make the management easier. An interval-valued symbolic random variableIn practice, the symbolic data is often generated by aggregating massive datasets into intervals in order to make the management easier. An interval-valued symbolic random variableIn practice, the symbolic data is often generated by aggregating massive datasets into intervals in order to make the management easier. An interval-valued symbolic random variableIn practice, the symbolic data is often generated by aggregating massive datasets into intervals in order to make the management easier. An interval-valued symbolic random variable

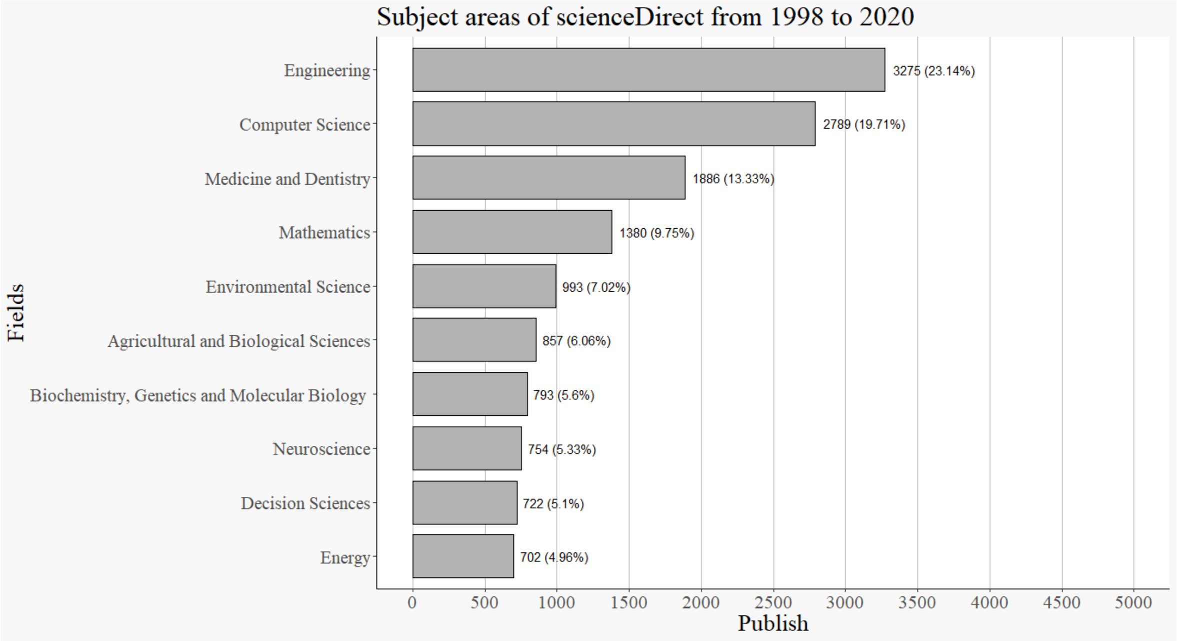

Figure 2: Top 10 researches and applications domains for SDA or interval-valued data (ScienceDirect) from 1998 to 2020

References

Billard L, Diday E (2007). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, New Jersey.

Diday, Edwin (2018). "New Advances on the Symbolic Data Analysis Framework: Basic Theory, Explanatory Criteria, Improving Machine Learning, New Directions of Research." p. 17.

Rojas OR (2015). "R to Symbolic Data Analysis." URL `https://www.imsbio.co.jp/RGM/R_rdfile?f=RSDA/man/RSDA-package.Rd&d=R_CC`.

Wickham H (2009). "**ggplot2**: Elegant Graphics for Data Analysis." *Media*, **35**(211), 10–1007. `doi:10.1007/978-0-387-98141-3`.

Affiliation:

Firstname Lastname
Affiliation
Address, Country
E-mail: name@address
URL: http://link/to/webpage/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd