Multivariate Analysis I

Alboukadel Kassambara

Practical Guide To Cluster Analysis in R

Unsupervised Machine Learning

sthda.com

Edition 1

Published by STHDA (http://www.sthda.com), Alboukadel Kassambara

Contact: Alboukadel Kassambara <alboukadel.kassambara@gmail.com>

For general information contact Alboukadel Kassambara <alboukadel.kassambara@gmail.com>.

0.1 Preface

Large amounts of data are collected every day from satellite images, bio-medical, security, marketing, web search, geo-spatial or other automatic equipment. Mining knowledge from these big data far exceeds human's abilities.

Clustering is one of the important data mining methods for discovering knowledge in multidimensional data. The goal of clustering is to identify pattern or groups of similar objects within a data set of interest.

In the litterature, it is referred as "pattern recognition" or "unsupervised machine learning" - "unsupervised" because we are not guided by a priori ideas of which variables or samples belong in which clusters. "Learning" because the machine algorithm "learns" how to cluster.

Cluster analysis is popular in many fields, including:

- In *cancer research* for classifying patients into subgroups according their gene expression profile. This can be useful for identifying the molecular profile of patients with good or bad prognostic, as well as for understanding the disease.
- In *marketing* for *market segmentation* by identifying subgroups of customers with similar profiles and who might be receptive to a particular form of advertising.
- In *City-planning* for identifying groups of houses according to their type, value and location.

This book provides a practical guide to unsupervised machine learning or cluster analysis using R software. Additionally, we developped an R package named *factoextra* to create, easily, a ggplot2-based elegant plots of cluster analysis results. Factoextra official online documentation: http://www.sthda.com/english/rpkgs/factoextra

0.2 About the author

Alboukadel Kassambara is a PhD in Bioinformatics and Cancer Biology. He works since many years on genomic data analysis and visualization. He created a bioinformatics tool named GenomicScape (www.genomicscape.com) which is an easy-to-use web tool for gene expression data analysis and visualization.

He developed also a website called STHDA (Statistical Tools for High-throughput Data Analysis, www.sthda.com/english), which contains many tutorials on data analysis and visualization using R software and packages.

He is the author of the R packages **survminer** (for analyzing and drawing survival curves), **ggcorrplot** (for drawing correlation matrix using ggplot2) and **factoextra** (to easily extract and visualize the results of multivariate analysis such PCA, CA, MCA and clustering). You can learn more about these packages at: http://www.sthda.com/english/wiki/r-packages

Recently, he published two books on data visualization:

1. Guide to Create Beautiful Graphics in R (at: https://goo.gl/vJ0OYb).
2. Complete Guide to 3D Plots in R (at: https://goo.gl/v5gwl0).

Contents

0.3 Key features of this book

Although there are several good books on unsupervised machine learning/clustering and related topics, we felt that many of them are either too high-level, theoretical or too advanced. Our goal was to write a practical guide to cluster analysis, elegant visualization and interpretation.

The main parts of the book include:

- *distance measures*,
- *partitioning clustering*,
- *hierarchical clustering*,
- *cluster validation methods*, as well as,
- *advanced clustering methods* such as fuzzy clustering, density-based clustering and model-based clustering.

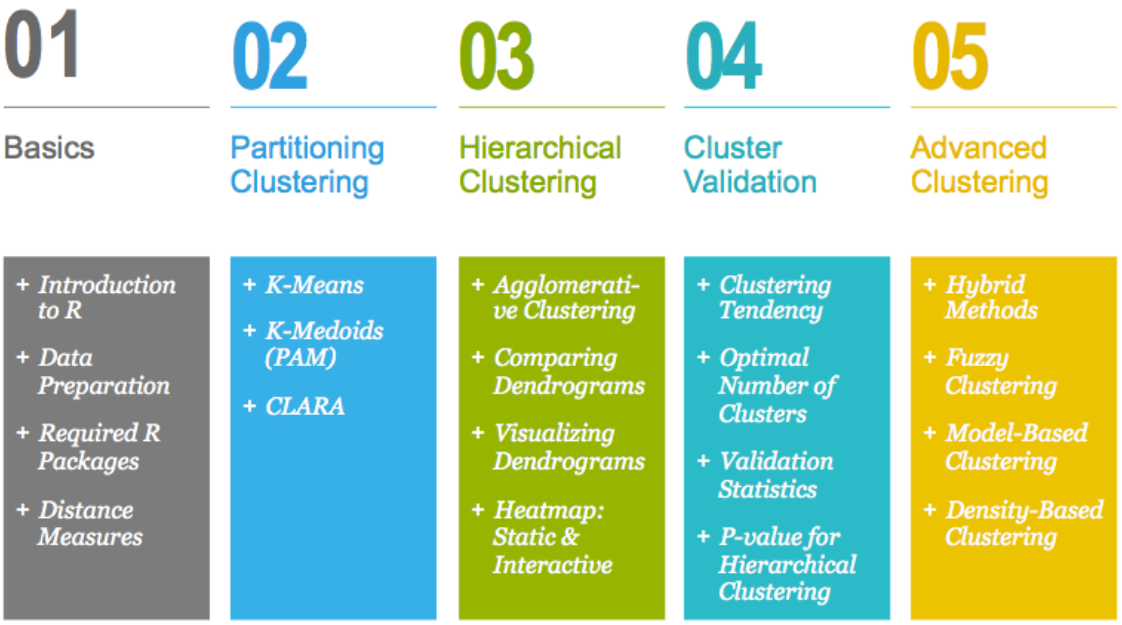
The book presents the basic principles of these tasks and provide many examples in R. This book offers solid guidance in data mining for students and researchers.

Key features:

- Covers clustering algorithm and implementation
- Key mathematical concepts are presented
- Short, self-contained chapters with practical examples. This means that, you don't need to read the different chapters in sequence.

At the end of each chapter, we present R lab sections in which we systematically work through applications of the various methods discussed in that chapter.

0.4 How this book is organized?



This book contains 5 parts. Part I (Chapter 1 - 3) provides a quick introduction to R (chapter 1) and presents required R packages and data format (Chapter 2) for clustering analysis and visualization.

The classification of objects, into clusters, requires some methods for measuring the distance or the (dis)similarity between the objects. Chapter 3 covers the common distance measures used for assessing similarity between observations.

Part II starts with partitioning clustering methods, which include:

- K-means clustering (Chapter 4),
- K-Medoids or PAM (partitioning around medoids) algorithm (Chapter 5) and
- CLARA algorithms (Chapter 6).

Partitioning clustering approaches subdivide the data sets into a set of k groups, where k is the number of groups pre-specified by the analyst.