R SDisc: Integrated methodology for data subtype discovery

Fabrice Colas*

March 14, 2011

Cluster analysis[1] is a statistical technique that aims to subset observations into groups, such that similar items are in the same clusters but are very different from items in other clusters. As a discovery tool, cluster analysis may enable to reveal associations, patterns, relationships, and structure in data. R SDisc is an additional tool to perform cluster analysis.

However, instead of proposing another clustering algorithm to the vast landscape of existing techniques [7], we focused on the development of a pipelined clustering analysis tool that would integrate the necessary tools and methods to run a complete analysis from data processing to subtype validation [1, 3]. It has been primarily designed for, and applied to clinical research on complex pathologies like Parkinson's disease [8], aggressive brain tumours [2] and Osteoarthritis where, more homogeneous patient subtypes from clinical predictors are sought for in order to break down the known clinical heterogeneity of those diseases (one disease-umbrella, different manifestations).

As such, R SDisc includes methods for data treatment and pre-processing, repeated cluster analysis, model selection, model reliability [4] and reproducibility assessment, subtype characterization and validation by visual and table summaries. In the design of R SDisc, we emphasized especially the validity of the inference steps, the accessibility of the cluster analysis protocol, the reproducibility of the results, and the availability as an open source package of the technique. This vignette is an interactive documentation on R SDisc.

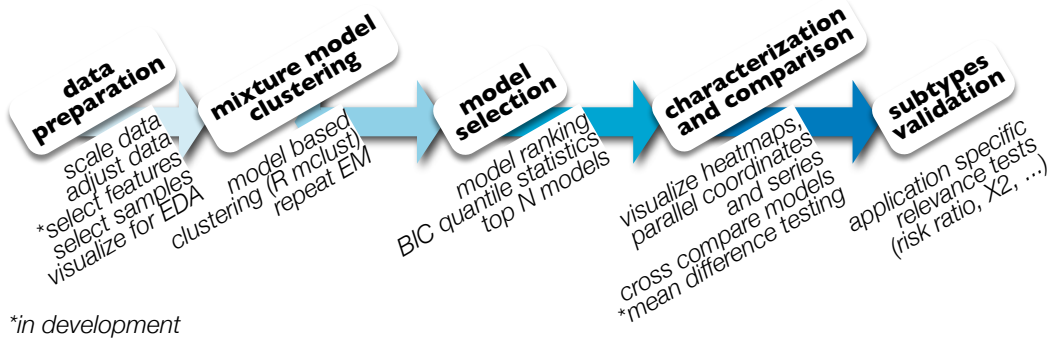

Figure 1: The data mining scenario consists in a sequence of five steps [3]: the data preparation, the cluster modeling based on [5, 6], the model selection, the characterization and comparison of the subtypes and the relevance evaluation.

*Leiden University Medical Center, Einthovenweg 20, 2300RC Leiden, the Netherlands

[1] http://www.businessdictionary.com/definition/cluster-analysis.html

Contents

1 R SDisc: Building blocks and Installation

```
> install.packages("SDisc", dep = TRUE)
```

```
> library(SDisc)
```

```
by using mclust, invoked on its own or through another package,
you accept the license agreement in the mclust LICENSE file
and at http://www.stat.washington.edu/mclust/license.txt
```

2 Data

R SDisc implements its own data structure to preserve a copy of the original data while applying some treatments to the data (feature selection, complete cases, adjustment), to limit cluster analysis to a subset of the predictors, to validate subtypes against clinical predictors not included in the cluster analysis (e.g. gender, bmi).

In this section, we introduce data examples that will be used throughout the hands on, to illustrate the functionnalities of R SDisc. Next, how R SDisc preprocess the data is introduced. Then, we present SDisc data container possibilities to do exploratory data analysis plots. In the case a previous processing was performed (e.g. year one of a longitudinal study), centering and adjustment estimates are retrieved and applied to the

new data. Last, when there are too many predictors (e.g. with spectral data), predictor selection by χ^2-testing can be performed.

2.1 Example

We simulate an example called `normdep.df` with three predictors; the first follows $\mathcal{N}(0, 1)$, the second is a time variable and the third depends on the time, provided $v = 2 \times t + \epsilon$. As a last step, we also inject 5 missing values at random into the data.

```
> set.seed(6015)
> Eps <- runif(50)
> time <- sample(1:5, 50, replace = TRUE)
> normdep.df <- matrix(c(rnorm(50), time, 2 * time + Eps), 50, 3,
      dimnames = list(list(), c("norm", "t", "v")))
> normdep.df[sample(1:nrow(normdep.df))[1:5]] <- NA
```

2.2 Preprocessing

```
> normdep.set <- SDDataSettings(normdep.df)
> normdep.set[, "tFun"] <- c("mean sd", "", "lm(v~t)")

> normdep <- SDData(normdep.df, settings = normdep.set, prefix = "normdep")

> SDDataSettings(normdep, latex = TRUE)
```

oddGroup inCAnalysis tFun vParGroup vParY vHeatmapY norm ”norm” ”TRUE” ”mean sd” ”varGroup1” ”1” ”1” t ”t” ”TRUE” ”” ”varGroup1” ”2” ”2” v ”v” ”TRUE” ”lm(v t)” ”varGroup1” ”3” ”3”

2.3 Exploration

```
> naPattern(normdep, latex = TRUE)
```

	isNA	isNotMissing	naRate
14	1.00	2.00	33.33
20	1.00	2.00	33.33
26	1.00	2.00	33.33
38	1.00	2.00	33.33
45	1.00	2.00	33.33

Table 1: normdep, index of the cases presenting **missing values** along with the number of missings and non-missings; the cases with a missing value represent 10.00% of the available cases.

```
> print(normdep, rseed = 6013, latex = TRUE)

> plot(normdep, latex = TRUE)

> summary(normdep, q = "mean|sd", latex = TRUE)

> summary(normdep, q = "lm", latex = TRUE)
```

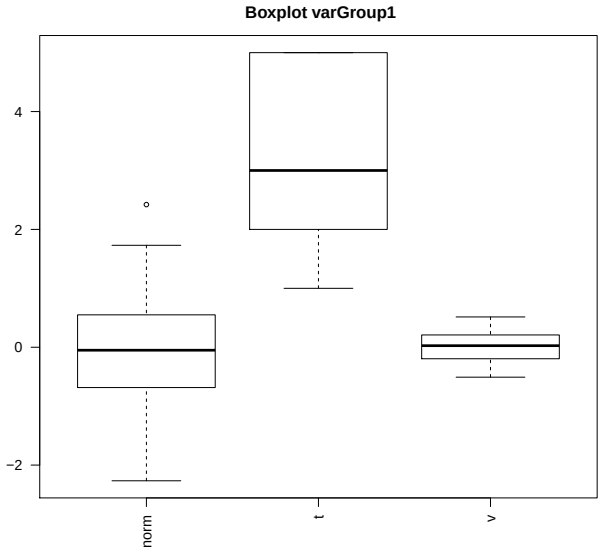

Figure 2: normdep, **boxplots** of the variables of the factor **varGroup1**.

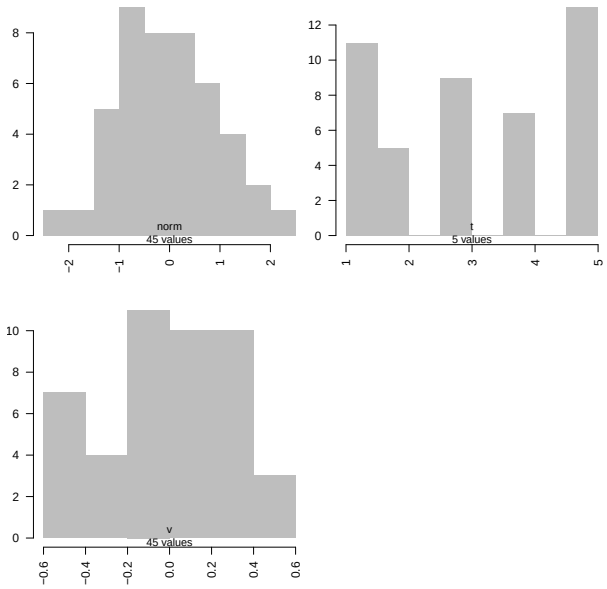

Figure 3: normdep, **histograms** of the variables of the factor **varGroup1**.

	t	v	norm
27	5.00	10.32	0.52
36	4.00	8.24	-1.57
21	2.00	4.12	-0.11

Table 2: normdep, extract of the **original** data matrix.

	t	v	norm
27	5.00	0.07	0.77
36	4.00	0.18	-1.66
21	2.00	0.36	0.04

Table 3: normdep, extract of the **transformed** data matrix.

	mean	sd
norm	-1.42e-01	8.63e-01

Table 4: normdep summary of the different data treatments operated on the data.

	(Intercept) (SE; Pr(>|t|))	t (SE; Pr(>|t|))	R^2 (adj-R^2; N)
v~t	0.55 (0.10; 2.0e-06)	1.97 (0.03; 1.5e-45)	0.99 (0.99; 45)

Table 5: normdep summary of the different data treatments operated on the data.

2.4 Processing Given Other Data Model

```
> set.seed(6016)
> Eps <- runif(30)
> time <- sample(1:5, 30, replace = TRUE)
> normdep.df2 <- matrix(c(rnorm(30), time, 2 * time + Eps), 30, 3,
      dimnames = list(list(), c("norm", "t", "v")))
> normdep2 <- predict(normdep, newdata = normdep.df2, prefix = "normdep2")

> summary(normdep2, q = "lm", latex = TRUE, sanitize = FALSE)
```

	(Intercept) (SE; Pr(>|t|))	t (SE; Pr(>|t|))	R^2 (adj-R^2; N)
v t	0.55 (0.10; 2.0e-06)	1.97 (0.03; 1.5e-45)	0.99 (0.99; 45)

Table 6: normdep2 summary of the different data treatments operated on the data.

```
> summary(normdep2, q = "mean|sd", latex = TRUE)
```

2.5 χ^2 Feature Selection of Spectral Data

todo

	mean	sd
norm	-1.42e-01	8.63e-01

Table 7: normdep2 summary of the different data treatments operated on the data.

3 Cluster Analysis

3.1 Repeated Model Based Cluster Analysis with MClust

```
> class(normdep)

[1] "SDData"

> normdep <- SDisc(normdep, settings = normdep.set, prefix = "normdep",
       cFunSettings = list(modelName = c("EII", "VII", "VEI", "VVI"),
           G = 3:5, rseed = 6013:6023))

 Prepare the data

 Load and test for consistency: normdep/IMAGE.RData

> class(normdep)

[1] "SDisc"

> class(SDData(normdep))

[1] "SDData"

> class(SDDataSettings(normdep))

[1] "matrix"

> normdep <- SDisc(normdep)
```

3.2 Likelihood-Based Model Selection

```
> summary(bicTable(normdep), latex = TRUE)
```

	EII	VII	VEI	VVI
3	13.54 (13.54, 13.55)	13.27 (13.27, 15.71)	0.00 (0.00, 4.11)	NA (6.99, 8.71)
4	14.40 (14.40, 14.78)	11.78 (11.78, 18.13)	5.77 (5.77, 6.43)	NA (8.71, 8.71)
5	13.13 (13.13, 14.83)	NA (10.26, 16.66)	4.75 (4.75, 12.32)	NA (NA, NA)

Table 8: normdep, model VEI,3,6020 shows the **highest BIC** score over: the repeated random starts, type of model and number of component.

```
> print(bicTable(normdep), modelName = "VII", G = 4, latex = TRUE)
```

	modelName	G	rseed	BIC	relativeBic
VII,4,6013	VII	4	6013	-377.88	11.78
VII,4,6016	VII	4	6016	-377.88	11.78
VII,4,6021	VII	4	6021	-377.88	11.78
VII,4,6014	VII	4	6014	-377.88	11.78
VII,4,6020	VII	4	6020	-377.88	11.78
VII,4,6023	VII	4	6023	-388.53	14.93
VII,4,6019	VII	4	6019	-388.53	14.93
VII,4,6015	VII	4	6015	-392.55	16.12
VII,4,6022	VII	4	6022	-393.63	16.44
VII,4,6018	VII	4	6018	-395.89	17.11
VII,4,6017	VII	4	6017	-400.50	18.47

Table 9: normdep, models whose **relative BIC** score difference is **less than 5%**.

	3	1	2
3			16
1		13	
2	16		

Table 10: normdep, the **comparison of model** VEI,3,6020 and VEI,3,6023 exhibits a random index 100.0 (a $\kappa = 100.0$, and a relative degree of association $V = 100.0\%$ with $p_{\chi^2} = 0.0005$, $\chi^2 = 90.0$).

4 Comparing, Characterizing, Validating and Testing Reproducibility of Subtypes

4.1 Cross Comparison Between Subtyping Analyses

```
> print(normdep, latex = TRUE)
```

```
> print(normdep, m1 = 1, m2 = bestModel(normdep, modelName = "VII",
      G = 4)[1], latex = TRUE)
```

You must provide one or two 'SDisc' objects, or two 'SDCModel'. NULL

4.2 Characterization of the Subtypes

```
> plot(normdep, latex = TRUE)
```

```
> summary(normdep, q = 1, latex = TRUE)
```

4.3 Validation of the Subtypes

todo with another dataset including a treatment group for instance

4.4 Reproducibility of Subtypes on New Data

todo use of predict.SDisc

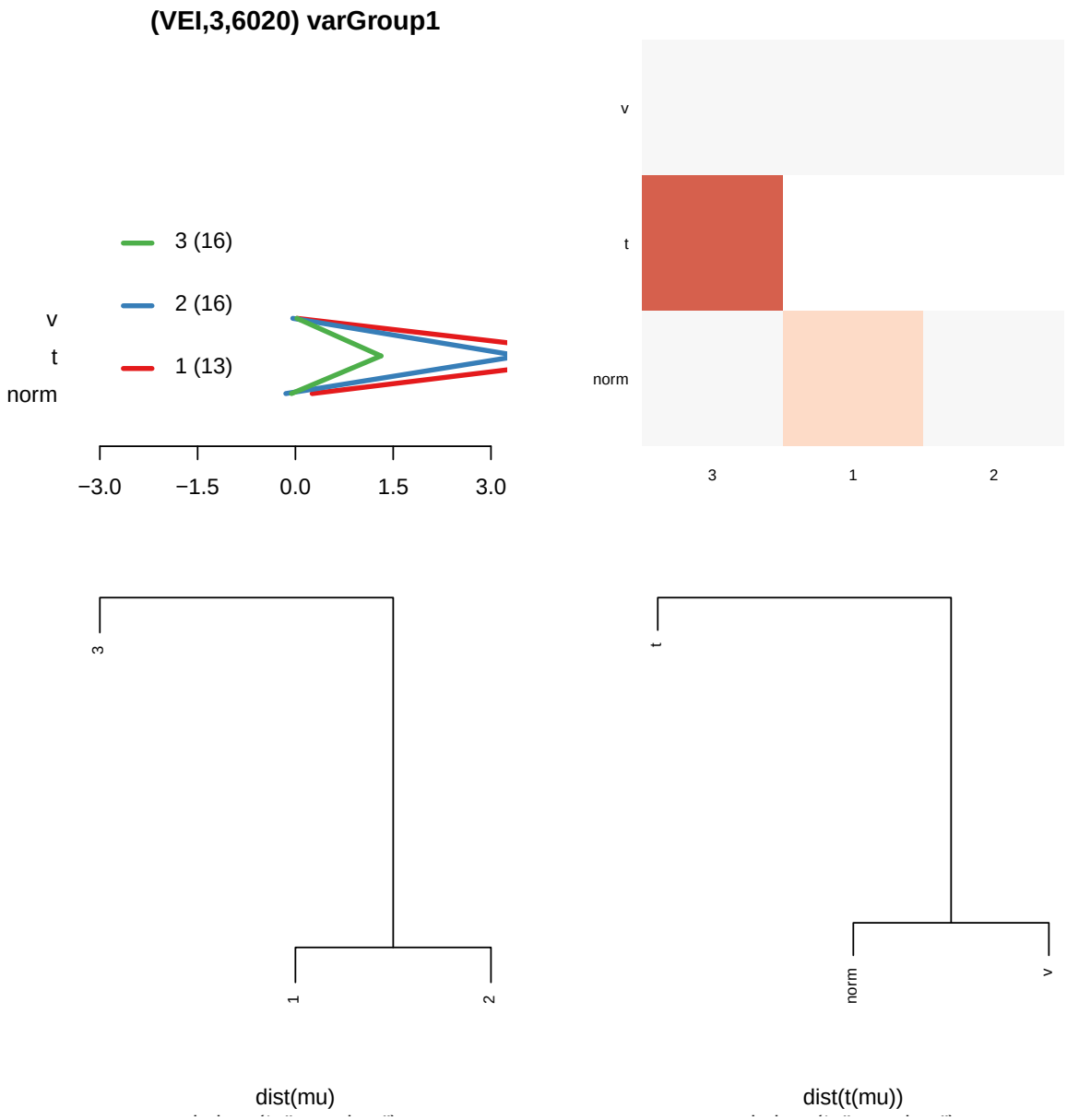

Figure 4: normdep, visual representation of **model** VEI,3,6020.

	1	2	3
norm	1.97	-0.44	-0.40
t	14.58	14.69	-15.33
v	0.37	-0.47	0.25

Table 11: normdep, (Bayesian) **oddratios** for the main factors in model VEI,3,6020.

A Session Info

```
> sessionInfo()

R version 2.12.1 (2010-12-16)
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)

locale:
[1] C/en_GB.UTF-8/C/C/en_GB.UTF-8/en_GB.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] SDisc_1.22          SparseM_0.86        snow_0.3-3          e1071_1.5-24
 [5] class_7.3-3         digest_0.4.2        xtable_1.5-6        abind_1.1-0
 [9] RColorBrewer_1.0-2 mclust_3.4.8

loaded via a namespace (and not attached):
[1] tools_2.12.1

> dir1 <- dir(recursive = TRUE)
> dir1 <- dir1[grep("pdf", dir1, invert = T)]
> file.remove(dir1[which(!(dir1 %in% dir0))])

logical(0)

> file.remove(c("normdep2/figures", "normdep2/tables"))

[1] TRUE TRUE

> file.remove(c("normdep2"))

[1] TRUE
```

List of Tables

List of Figures

References

[1] Fabrice Colas. *Data Mining Scenarios for the Discovery of Subtypes and the Comparison of Algorithms*. phd, Leiden University, 2009.

[2] Fabrice Colas, Joost N. Kok, and Alfredo Vellido. Finding discriminative subtypes of aggressive brain tumours using magnetic resonance spectroscopy. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 1:1065–8, 2010 2010.

[3] Fabrice Colas, Ingrid Meulenbelt, Jeanine J. Houwing-Duistermaat, Margreet Kloppenburg, Iain Watt, Stephanie M. van Rooden, Martine Visser, Johan Marinus, Edward O. Cannon, Andreas Bender, Jacobus J. van Hilten, P. Eline Slagboom, and Joost N. Kok. A scenario implementation in r for subtypediscovery examplified on chemoinformatics data. In *Leveraging Applications of Formal Methods, Verification and Validation, Communications in Computer and Information Science*, volume 17, pages 669–683. Springer Berlin Heidelberg, Springer Berlin Heidelberg, 2008.

[4] Fabrice Colas, Ingrid Meulenbelt, Jeanine J. Houwing-Duistermaat, Margreet Kloppenburg, Iain Watt, Stephanie M. van Rooden, Martine Visser, Johan Marinus, Jacobus J. van Hilten, P Slagboom, and Joost N. Kok. Reliability of cluster results for different types of time adjustments in complex disease research. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2008:4601–4, 2008 2008.

[5] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[6] C. Fraley and A. E. Raftery. MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics, September 2006.

[7] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31:264–323, September 1999.

[8] Stephanie M. van Rooden, Fabrice Colas, Pablo Martínez-Martín, Martine Visser, Dagmar Verbaan, Johan Marinus, Ray K Chaudhuri, Joost N. Kok, and Jacobus J. van Hilten. Clinical subtypes of parkinson's disease. *Movement disorders : official journal of the Movement Disorder Society*, 2010 Nov 16 2010.