R SubtypeDiscovery: a data mining scenario for the inference of subtypes by cluster analysis

F. Colas

August 3, 2009

Introduction

In the study of phenomena characterized by heterogeneity, an important and general data analysis problem is the search for more homogeneous subtypes in the data distribution. In clinical research such as on Osteoarthritis [Meul97], Parkinson's disease [Roo09], major depressive and anxiety disorders [Cla91], or glioblastoma and metastasis discrimination [Vel09], the identification of more homogeneous patient subtypes may contribute to understand more specifically the underlying mechanisms of these pathologies, and thus help to develop tailored prevention strategies and treatments.

To advance research on these phenomena, we developed a data-mining scenario designed to infer subtypes by cluster analysis [Col09b, Col09a]. This scenario is referred to as SubtypeDiscovery and it was implemented as an R package. As a result, other research teams can benefit of this tested scenario to perform their subtyping analyses. With this package, analyses are straightforward and therefore accessible to many investigators. Furthermore, the well-defined data structures and the public availability of the package greatly enhance reproducibility.

The scenario features various data preparation techniques, an approach that repeats data modeling in order to select for the number of subtypes and/or the type of model, with a selection of methods to characterize, compare and evaluate the most likely mixture models. Table 1 describes the different steps of the data analysis.

Table 1: Combination of steps of a typical subtype discovery analysis.

1. data configuration, processing and exploratory data analysis (EDA),
2. model based clustering [Fra02, Fra06] repeated from different initialization points,
3. selection of the top models based on a Bayesian Information Criterion ranking,
4. cross-comparison of those mixture models and characterization of their subtypes,
5. relevance evaluation of each subtype.

Install SubtypeDiscovery, and set up the R environment

The website of SubtypeDiscovery is

`https://gforge.nbic.nl/projects/subtypediscover/`

To download the package, follow the *download* link and retrieve the most recent one; for windows, choose the zip file, otherwise the tar.gz. Next, install the package via the menu on windows or via the command line on a Unix (`R CMD install SubtypeDiscovery_1.16.[zip/tar.gz]`). Then, load in your actual R environment the package via the command

```
> library(SubtypeDiscovery)
```

```
by using mclust, you accept the license agreement in the LICENSE file
and at http://www.stat.washington.edu/mclust/license.txt
```

> Note that the use of the model based clustering library (`mclust`) is subject to a licence agreement and that, except for strict academic use, a licence fee must be paid, see `licence.txt`. The licence also requires to refer to a technical report and an article [Fra02, Fra06].

As SubtypeDiscovery depends on a number of packages (`mclust, stats, utils, RColorBrewer, abind, e1071, xtable`), it is likely that some are missing in your R environment or that some versions are unmet. To install the ones missing, do

```
> install.packages('mclust','stats','utils','RColorBrewer','abind','e1071','xtable')
```

Prepare the data for the analysis (cdata)

In the package, a public chemoinformatics dataset, contributed by Ed O Cannon, is included (see the man page `?wada2008`). We use it to perform a sample subtype discovery analysis illustrating the various features of the package. Meanwhile, advices are provided on how to interpret the results and how to carry the inference of subtype.

First, we load the chemoinformatics dataset.

```
> data(wada2008)
```

Second, we define a `settings` file describing how SubtypeDiscovery can interpret the data. We save the file as a `csv` to edit it outside of R. As such, a more appropriate ordering of the variables may be specified, certain variables can be discarded from the cluster analysis, etc.

```
> fSettings <- genCdataSettings(wada2008, asCSV = TRUE)
> print(fSettings)
```

```
[1] "2009-08-03_settings.csv"
```

Using Excel/Openoffice, we edit the settings, save them and read them back into R.

```
> settings <- read.csv(fSettings, row.names = 1)
> colnames(settings)
```

```
[1] "oddGroup"    "inCAnalysis" "tFun"        "vParGroup"   "vParY"
[6] "vHeatmapY"
```

The column of a settings file are as follows. `oddGroup` defines sum factors to calculate odd ratios from the factor distribution, for each subtype. `inCAnalysis` (TRUE/FALSE) tells whether the variable is included for the cluster analysis. `tFun` defines the series of transformations to apply to the variables. `vParGroup` gives the factors for graphical characterization of the data by parallel coordinates, whereas `vParY` is the y-coordinate of the parallel coordinate axis. Last, `vHeatmapY` describes the ordering of the variable in the heatmap graphical characterization.

However, for the purpose of this sample analysis, we already edited settings for `wada2008`. They are referred to as `wada2008_settings` and they are saved within the SubtypeDiscovery package installation. To load `wada2008_settings` into the environment, do

```
> data(wada2008_settings)
> wada2008_settings[c(50:54), ]
```

```
            oddGroup                inCAnalysis tFun
VAdjMa      "Atom and bond counts" "TRUE"      "tAvg tSigma"
VDistEq     "Atom and bond counts" "TRUE"      "tAvg tSigma"
VDistMa     "Atom and bond counts" "TRUE"      "tAvg tSigma"
Data.Source NA                     "FALSE"     NA
Labels      NA                     "FALSE"     NA
            vParGroup                  vParY   vHeatmapY
VAdjMa      "Atom and bond counts (2)" " 1.45" "50"
VDistEq     "Atom and bond counts (2)" " 1.00" "51"
VDistMa     "Atom and bond counts (2)" " 0.55" "52"
Data.Source NA                         NA      NA
Labels      NA                         NA      NA
```

Thus, we can proceed to the instantiation of the dataset on which the sample SubtypeDiscovery analysis will be performed.

```
> wada2008_cdata <- setCdata(data = wada2008[1:100, ], settings = wada2008_settings,
+     prefix = "Wada2008")
```

`setCdata` automatically backs-up the data in a dedicated space, it processes each variable according to the transformation specified in `settings`, it saves the transformation results (e.g. the estimated mean or standard deviation) and returns the processed data as a `cdata` object, which contains everything.

We can summary a `cdata` using the `summary()` function, it results in

```
> summary(wada2008_cdata)
```

```
 ' Wada2008 ' dataset for subtype discovery analysis
 R version 2.8.1 (2008-12-22) , i386-apple-darwin8.11.1
 SubtypeDiscovery  1.16
          number of variables in the cluster analysis/originally:  98 / 101
          number of complete and incomplete cases:  100 / 0
```

```
xlim:  -3 3
ylim:  0 50
par() 'mai' parameter:  0.6 0.3 0.05 0.05 (to define the base margin of the pl
base, figure and table directories:
       2009-08-03_Wada2008-b
       2009-08-03_Wada2008-b/figures
       2009-08-03_Wada2008-b/tables
```

The use of the `print()` function on a `cdata` object will return the data matrix after all the data processings.

```
> print(wada2008_cdata)[1:5, 1:5]
```

```
      balabanJ  diameter    KierFlex   petitjean petitjeanSC
1  0.658827831 0.7862488 -0.01368051 -0.18609105  0.03073845
2 -0.003183211 0.2789915 -0.30777615 -0.22560478 -0.02158100
3 -0.601845714 1.0398774 -0.21020500  0.03123497  0.34465462
4  0.477695892 0.7862488  0.53984687 -0.18609105  0.03073845
5 -1.927577651 0.5326201  0.22669357  0.03123497  0.34465462
```

A `cdata` object also has a number of graphical methods, which may be used for exploratory data analysis of the data (EDA). Thus, using the `plot()` function results in

```
> plot(wada2008_cdata)
```

```
[1] "2009-08-03_Wada2008-b/figures/oddGroup_001-BB.pdf"
[1] "2009-08-03_Wada2008-b/figures/oddGroup_001-H.pdf"
[1] "2009-08-03_Wada2008-b/figures/oddGroup_002-BB.pdf"
[1] "2009-08-03_Wada2008-b/figures/oddGroup_002-H.pdf"
[1] "2009-08-03_Wada2008-b/figures/oddGroup_003-BB.pdf"
[1] "2009-08-03_Wada2008-b/figures/oddGroup_003-H.pdf"
[1] "2009-08-03_Wada2008-b/figures/oddGroup_004-BB.pdf"
[1] "2009-08-03_Wada2008-b/figures/oddGroup_004-H.pdf"
[1] "2009-08-03_Wada2008-b/figures/oddGroup_005-BB.pdf"
[1] "2009-08-03_Wada2008-b/figures/oddGroup_005-H.pdf"
```

This EDA produces a number of figures describing the data, e.g. boxplots (`BB`) and histograms (`H`). The index of these figures refer to the visualization groups defined in the `settings` file. In Figure 1, we illustrate a boxplot and a series of histograms for the first group (`adjency, distance matrix, kier and hall, connectivity, kappa shape indices`).

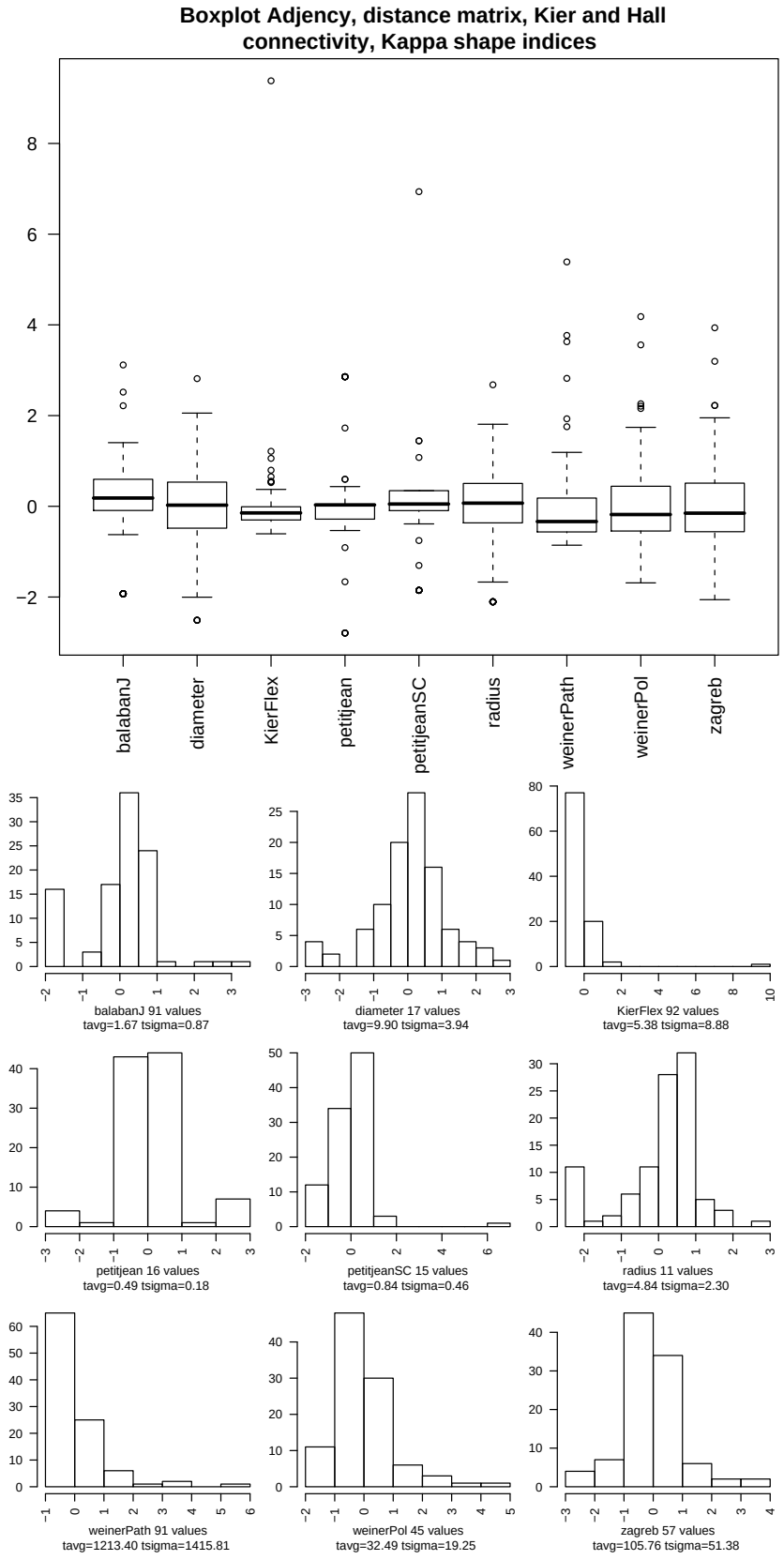

Figure 1: These graphics summarize the distributional properties of the dataset variables, which may enable to identify the presence of outlying values, to screen the existence of variables that are binary or that show little continuity. This may lead to the exclusion of some variables, of cases, or to the choice of an alternative data processing.

SubtypeDiscovery analysis, set-up and calculation

In the previous section, we presented the steps to prepare a dataset. In this, we show how to perform the analysis itself.

The dataset `cdata` is the only mandatory parameters of `setCresult`, the others have default values for a simple analysis.

> `setCresult` does accept a number of additional parameters to adapt the calculations to the application area. Among these parameters, there is first the cluster modeling method (`cfun`), whose parameters are provided in `cfun_params`. There is, too, the graphic characterization of the mixture models, which is defined by a `fun_plot` parameter expecting a list of functions provided by `getPlotFun`. With a similar mechanism, a number of statistical methods to characterize or evaluate the results of an analysis can be defined into the `fun_stats` parameter, which expects a list of function that are result of `getStatsFun`. Finally, the parameter `nTopModels` specifies how many top-ranking models will be selected as likely and, thus, be cross-compared for consistency assessment. More details are provided in the man page of `?setCresult`.

Thus, to instantiate the SubtypeDiscovery analysis and to start the calculations, on `wada2008`, we do

```
> x <- setCresult(cdata = wada2008_cdata)
> x <- doModeling(x)

 EII,3,6013-> Patterns-> Dendros-> Ordering-> Stats
 VII,3,6013-> Patterns-> Dendros-> Ordering-> Stats
 EII,4,6013-> Patterns-> Dendros-> Ordering-> Stats
 VII,4,6013-> Patterns-> Dendros-> Ordering-> Stats
 EII,5,6013-> Patterns-> Dendros-> Ordering-> Stats
 VII,5,6013-> Patterns-> Dendros-> Ordering-> Stats
 EII,3,6014-> Patterns-> Dendros-> Ordering-> Stats
 VII,3,6014-> Patterns-> Dendros-> Ordering-> Stats
 EII,4,6014-> Patterns-> Dendros-> Ordering-> Stats
 VII,4,6014-> Patterns-> Dendros-> Ordering-> Stats
 EII,5,6014-> Patterns-> Dendros-> Ordering-> Stats
 VII,5,6014-> Patterns-> Dendros-> Ordering-> Stats
 EII,3,6015-> Patterns-> Dendros-> Ordering-> Stats
 VII,3,6015-> Patterns-> Dendros-> Ordering-> Stats
 EII,4,6015-> Patterns-> Dendros-> Ordering-> Stats
 VII,4,6015-> Patterns-> Dendros-> Ordering-> Stats
 EII,5,6015-> Patterns-> Dendros-> Ordering-> Stats
 VII,5,6015-> Patterns-> Dendros-> Ordering-> Stats
Save modeling into  2009-08-03_Wada2008-b/IMAGE.RData

> summary(x)

' Wada2008 ' subtype discovery analysis summary
------- data ----------
  ' Wada2008 ' dataset for subtype discovery analysis
```

```
  R version 2.8.1 (2008-12-22) , i386-apple-darwin8.11.1
  SubtypeDiscovery  1.16
           number of variables in the cluster analysis/originally:  98 / 101
           number of complete and incomplete cases:  100 / 0
           xlim:  -3 3
           ylim:  0 50
           par() 'mai' parameter:  0.6 0.3 0.05 0.05 (to define the base margin of the p
           base, figure and table directories:
                 2009-08-03_Wada2008-b
                 2009-08-03_Wada2008-b/figures
                 2009-08-03_Wada2008-b/tables
----- settings --------
  R version 2.8.1 (2008-12-22) , i386-apple-darwin8.11.1
 SubtypeDiscovery  1.16 )
      model based cluster analysis
            covariance models:  EII VII
            number of clusters:  3 4 5
            random initialization numbers:  6013 6014 6015
            BIC table of relative difference with respect to most likely model
  EII                    VII
3 "20.52 (20.52, 20.52)" "7.95 (7.95, 7.95)"
4 "17.49 (17.49, 17.49)" "7.91 (7.91, 7.91)"
5 "18.05 (18.05, 18.05)" "1.73 (1.73, 1.73)"
  EII                                 VII
3 "-23055.64 (-23055.64, -23055.64)" "-20650.41 (-20650.41, -20650.41)"
4 "-22476.07 (-22476.07, -22476.07)" "-20643.88 (-20643.88, -20643.88)"
5 "-22582.73 (-22582.73, -22582.73)" "-19460.73 (-19460.73, -19460.73)"
  EII              VII
3 "1 (1.00, 1.00)" "1 (1.00, 1.00)"
4 "3 (3.00, 3.00)" "2 (2.00, 2.00)"
5 "2 (2.00, 2.00)" "3 (3.00, 3.00)"
      model ranking, top:
                 1   VII,5,6015
                 2   VII,5,6014
                 3   VII,5,6013
                 4   VII,4,6014
                 5   VII,4,6015
```

The output of `doModeling` describes the sequence of calculations performed. First, there is the calculation of the mixture models of type (`EII,VII`), of number of components `3,4,5`, and of random initialization integers `6013,6014,6015`. Then, statistical patterns such as the empirical mean, the standard deviation, or other quantile statistics are estimated during the `Patterns` step. For each mixture model, by default, a euclidean distance based hierarchical clustering is performed on both the variables and the observations (`Dendros`), and then, it will be used reorder the patterns (`Ordering`). Finally, statistics for each subtype are calculated during the `Stats` step, e.g. edds ratios, which rely on the `oddGroup` column of the `settings`. For more details, see [Col09a].

Once calculation is completed, the subtype discovery analysis (`cresult`) is stored on the hard drive into an `RData` file (`IMAGE.RData`). Further, to enable post-hoc analysis of the

discovered subtypes, the set of best mixture models is stored into `csv` files archived under the `tables/` directory. These files report the likelihood of every element to belong to each mixture component, along with the component affectation.

```
> plot(x)

[1] "2009-08-03_Wada2008-b/figures/MM-EII_3_6013.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-VII_3_6013.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-EII_4_6013.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-VII_4_6013.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-EII_5_6013.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-VII_5_6013.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-EII_3_6014.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-VII_3_6014.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-EII_4_6014.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-VII_4_6014.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-EII_5_6014.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-VII_5_6014.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-EII_3_6015.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-VII_3_6015.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-EII_4_6015.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-VII_4_6015.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-EII_5_6015.pdf"
[1] "2009-08-03_Wada2008-b/figures/MM-VII_5_6015.pdf"
```

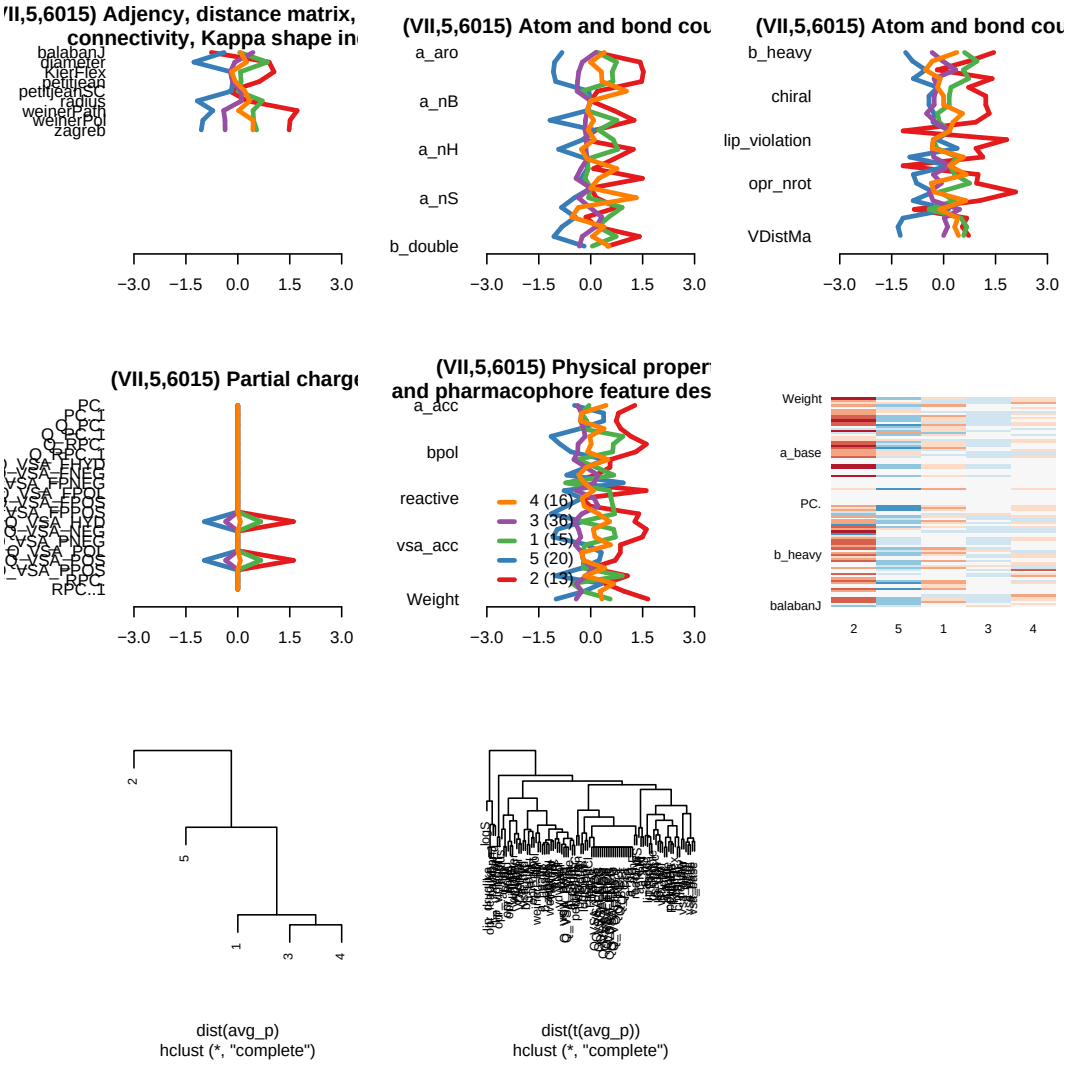

Figure 2: Graphical characterization of a `cresult`.

Carrying the statistical inference

In previous two sections, we first presented how to set-up a dataset for a SubtypeDiscovery analysis and then, we showed how to perform a small sample analysis. In the following, to identify data subtypes in the reduced `wada2008` dataset, we propose a path to carry the inference. For this purpose, we will use a number of graphics and summary measures.

First, we report a table of likelihood-based scores enabling the comparison of mixture models whose number of parameters is different. This score is referred to as the Bayesian Information Criterion and it calculates a trade-off between model likelihood and number of parameters. A trade-off measure is required because the more the number of parameters of a model (number of mixture components, of variance and covariance parameters, etc.), the more the model will fit the data distribution. Yet, because of its many parameters, the model may not show good generalization properties, a phenomena known as overfitting.

To get this table of BIC scores, we do

```
> bAn <- attr(x, "bicanalysis")
> bAn.avg <- summary(bAn, type = "latex", fun = "mean", caption = paste2("Average rela
```

	EII	VII
3	20.52 (20.52, 20.52)	7.95 (7.95, 7.95)
4	17.49 (17.49, 17.49)	7.91 (7.91, 7.91)
5	18.05 (18.05, 18.05)	1.73 (1.73, 1.73)

Table 2: Average relative BIC score difference with respect to the top ranking model.

As reported by the Table 2, the model having the highest likelihood in terms of BIC score occurs for 5 subtypes, a covariance model of type VII, and an initialization number 6015. Further, we also read that models (VII,5,*) show, in average, a relative difference of 1.73 (1.73, 1.73)% with respect to VII,5,6015. Similarly, models (EII, 4,*) exhibit an average relative BIC score difference of 20.52 (20.52, 20.52)%.

> Yet, in some analysis, the observed BIC score differences are relatively smalls, e.g. less than 5%, which is not necessarily significant. Our strategy is therefore to filter out the models having an average difference greater than 5% and to focus on those whose average BIC is less than 5%. Alternatively, the top ranking models can be considered.

In the present analysis, the top 5 is considered and the models are

```
> getBestModel(x)

[1] "VII,5,6015" "VII,5,6014" "VII,5,6013" "VII,4,6014" "VII,4,6015"
```

Second, to compare two models, e.g. the two best ones in terms of BIC scores. We use the `print` function, we apply it to the `cresult` object, and we provide it the name of the two first mixture model names:

```
> print(x, m1 = getBestModel(x, 1), m2 = getBestModel(x, 2), type = "latex")
```

	3	1	4	2	5
3	30				6
1		14	1		
4		15			1
2			11	2	
5				11	9

Table 3: The level of association of models VII,5,6015 and VII,5,6014 is $V = 76.8\%$ and the χ^2 test, on which V is based, has its $p = 0.0005$ ($\chi^2 = 235.9$).

Table 3 reports the joint distribution of the cluster affectation of the two mixture models. The joint distribution enables to report the level of consistency of two models. If most of the table elements are in the diagonal, and there are many 'empty' cells, there is very good association between the two models. As well, when comparing two mixture models of the same type, for a growing number of mixture components, one may assess whether there is a nested structure in the subtype.

To summary that joint distribution and thus, the level of association between two cluster models, we use the Cramer's V. Similarly to Pearson's correlation coefficient, the Cramer's V takes values in [0, 1], where one stands for completely correlated variables and zero for stochastically independent ones. The measure is symmetric and it is based on the χ^2 statistics of nominal association. Therefore, the more unequal the marginals, the more V will be less than one. Alternatively, the measure can be regarded as a percentage of the maximum possible variation between two variables.

In Table 3, the Cramer's V is equal to $V = 76.79\%$.

In some application domains, it is possible to group variables by main factor, e.g. the main joint sites in Osteoarthritis (the spine facets, the spine lumbars, the hips, the knees, the distal and the proximal interphalengeal joints), the impairment domain in Parkinson's disease (the cognitive, the motricity and the autonomic disorders) and the class of molecular descriptors, in chemoinformatics.

To characterize the subtypes on each of these main factors, we compute the odd of the subtype distribution as compared to the one of the dataset. Odd ratios may exhibit significant subtype-specific distributional disparities.

In Table 4, we report the odd ratios for the most likely model VII,5,6015.

```
> summary(x[[getBestModel(x, 1)]], type = "latex")
```

More result interpretation in future versions of the vignette...

	1	2	3	4	5
A.D.M.D.	Inf	1.91	-2.14	Inf	-2.09
A.B.C.	Inf	Inf	-2.14	Inf	-Inf
K.H.C.K.S.I.	Inf	Inf	-2.69	Inf	-2.60
P.C.D.	Inf	1.91	-1.90	1.71	-2.09
P.F.D.	0.81	Inf	-2.69	Inf	-0.77
P.P.	Inf	Inf	-1.90	1.71	-Inf

Table 4: Statistics of model VII,5,6015 (oddratios).

Concluding remarks

In this vignette, we presented the SubtypeDiscovery data mining scenario to infer subtypes in data, along with implementation as an R package. First, we described the package installation procedure, indicating the different packages it relied on. Then, we presented the dataset preparation procedure and we illustrated its principal functions. Similarly to the coverage of the dataset preparation, we described how to set-up a simple subtyping analysis, how to do the calculations, and how to get a number of pivotal tables and graphics.

References

[Cla91] Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. Lee Anna Clark and David Watson, Journal of Abnormal Psychology, 100 (3), pp 316-336, 1991.

[Col08] Stability of Clusters for Different Time Adjustments in Complex Disease Research. F Colas and I Meulenbelt and J J Houwing-Duistermaat and M Kloppenburg and I Watt and S M van Rooden and M Visser and H Marinus and J J van Hilten and P E Slagboom and J N Kok, 30th Annual International IEEE EMBS Conference (EMBC'08), Vancouver, Canada

[Col09a] A Scenario Implementation in R for Subtype Discovery Examplified on Chemoinformatics Data. F Colas and I Meulenbelt and J J Houwing-Duistermaat and M Kloppenburg and I Watt and S M van Rooden and M Visser and H Marinus and Edward O Cannon and Andreas Bender and J J van Hilten and P E Slagboom and J N Kok, 3rd International Symposium on Leveraging Applications of Formal Methods, Verification and Validation (ISOLA'08), Greece.

[Col09b] R SubtypeDiscovery to assist subtyping analyses in clinical research. F Colas, S van Rooden, I Meulenbelt, J J Houwing-Duistermaat, T van Veen and J N Kok in the Bioinformatics journal, Vol X, Number X, XXX 2009.

[Fra02] Model-Based Clustering, Discriminant Analysis and Density Estimation. C Fraley and A E Raftery. In Journal of the American Statistical Association, Vol 97, 611-631, 2002.

[Fra06] MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. C Fraley and A E Raftery. In technical report 504, Department of Statistics, University of Washington, September, 2006.

[Meul97] Genetic predisposing factors of osteoarthritis. Ingrid Meulenbelt, PhD thesis Leiden Universiteit, 1997.

[Roo09] Subtypes in Parkinson's disease. Stephanie M van Rooden and Fabrice Colas and Pablo Martinez-Martín and Martine Visser and D. Verbaan and Johan Marinus and Kallol Ray Chaudhuri and Joost N. Kok and Jacobus J van Hilten, submitted for publication, 2009.

[Vel09] Outlier exploration and diagnostic classification of a multi-centre 1H-MRS brain tumour database. Alfredo Vellido, Enrique Romero, Félix F. González-Navarro, Lluís A. Belanche-Muñoz, Margarida Juliá-Sapé, Carles Arús (2009) Neurocomputing, 72(13-15), 3085-3097.