An R Package for Subtype Discovery Examplified on Chemoinformatics Data

Fabrice Colas[1], Stephanie M. van Rooden[2], Ingrid Meulenbelt[3], Jeanine J. Houwing-Duistermaat[4], Andreas Bender[5], Edward O. Cannon[6], Martine Visser[2], Johan Marinus[2], Jacobus J. van Hilten[2], P. Eline Slagboom[3], and Joost N. Kok[1,3]

[1] (fcolas@liacs.nl) LIACS, Leiden University, The Netherlands
[2] Neurology Department, Leiden University Medical Center, The Netherlands
[3] MOLEPI, Leiden University Medical Center, The Netherlands
[4] MEDSTATS, Leiden University Medical Center, The Netherlands
[5] LACDR, Leiden University, The Netherlands
[6] UCMSI, University of Cambridge, United Kingdom

Abstract. We developed a methodology that both facilitates and enhance the search for homogeneous subtypes in data. We applied this methodology to medical research on Osteoarthritis and Parkinson's Disease and to chemical databases in chemoinformatics. We release this methodology as the R SubtypeDiscovery package to enable *reproducibility* of our analyses. In this paper, we present the package implementation and we illustrate its output on molecular data from chemoinformatics, which we bring public. Our methodology includes different techniques to process the data, a computational approach repeating data modelling to select for a number of subtypes or a type of model, and additional methods to characterize, compare and evaluate the top ranking models. Therefore, this methodology does not solely cluster data but it also produces a complete set of results to conduct a subtype discovery analysis.

1 Introduction

In medical research, it is of interest to identify subtypes of diseases like Osteoarthritis (OA) and Parkinson's Disease (PD) that present clinical heterogeneity. We can do so by searching for homogeneous clusters in values of markers that reflect the severity of the disease. For chemoinformatics, in order to understand the relationship between different bioactivity classes of molecules, subtype discovery of chemical databases may improve our understanding of the similarity (and distance) between different phenotypic effects as induced by drugs and chemicals.

To this aim, we developed a methodology mimicking a cluster analysis process: from data preparation to cluster evaluation. In particular, it implements various data preparation techniques to facilitate the analysis given different data processing [1]. It also features a computational approach that repeats data modelling in order to select for a number of subtypes or a type of model. Additionally,

it defines a selection of methods to characterize, compare and evaluate the top ranking models [2].

As the search for subtypes appears in many areas, we abstract from the application we have done up to now and make it available as the R SubtypeDiscovery package. We present its implementation in this paper. The outline is as follows: in section 2, we give the main features of our methodology, in section 3 we discuss the design of the package, which we illustrate by an example on the chemoinformatics molecular data.

2 A methodology for subtype discovery

The methodology consists of a number of steps. We next discuss these steps in detail.

Data preparation. As data preparation can influence largely the result of data analyses, our methodology includes various methods to transform and process data, e.g. computing the z-scores of variables to obtain scale-invariant quantities. Alternatively, we may want to remove the time dimension in the data because we do not want to model clusters only characterized by the time. As an example, age in OA and disease duration in PD are two time variables known to play a major role in the overall severity of these diseases. To proceed, we analyze the residual variance of a regression on the time [1]. Other methods implement L_2/L_1 and max data normalization, and centering with respect to the $mean$, the $median$ or the min.

Cluster analyses. We use the model based clustering framework developed by Fraley and Raftery [3]. As shown in [4], the framework relies on the concept of reparameterization of the covariance matrix which enables to select and adapt the level of complexity of the covariance by controlling its geometry, see Fig. 1.

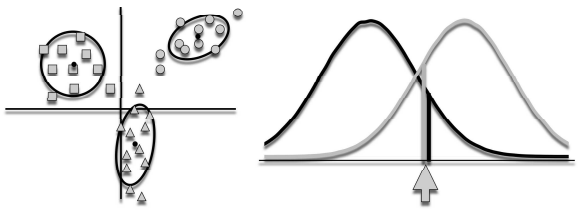

Fig. 1. On the left, we illustrate a simple modelling with three mixtures in two dimensions which are defined by their center μ_k and their geometry Σ_k, $k = 1, 2, 3$. On the right, we illustrate two mixtures on a single dimension. The gray is most likely and determines the cluster membership. The black is less likely and informs on the clustering uncertainty.

For a given number of mixtures and a covariance model, the EM-algorithm is used to estimate the model parameters. It alternates iteratively between Expectation to estimate for each observation its cluster membership likelihood, and Maximization to optimize the model parameters that maximize the likelihood. Then the iterative process stops as likelihood improvements become very small. Moreover, as the starting point of EM may influence the final result, in our analyses we repeat model estimation given different starting points. We then use the starting point that leads to the most likely model.

BIC analysis. The larger the number of parameters, the more likely it is that our model may overfit the data which restricts its generality and comprehensiveness. Therefore, to select the most likely model, Kass and Raftery [5] prefer the Bayesian Information Criterion (BIC) to the Akaike Information Criterion (AIC) because it approximates the Bayes Factor; we use the BIC in our analyses, $BIC = -2\log\mathcal{L}_{MIX} + \log(N \times \sharp params)$. We further approach the problem of selecting a number of subtypes and a type of model computationally by repeating the data modeling. Thus, especially analyzing the BIC scores of those models, we report in first place a BIC table that aggregates the best scores given all repeats. Second, we provide rankings on models, number of clusters and starting values. Finally, in another set of tables, we characterize those BIC scores given their mean, standard deviation, median, 2.5 and 97.5% quantiles. See Table 2 for an extract output of those tables.

Selected methods to characterize, compare and evaluate subtypes. To more easily evaluate the influence on the cluster results of different data preparation or to compare two by two cluster results, we need efficient visualization tools to see the prominent characteristics of the cluster results. Influenced by Tukey [6] and Tufte [7,8] for scientific data visualization and by Brewer's suggestions for color selection in geography [9], we selected three types of visual-aids, namely the heatmaps [10], the dendrograms of hierarchical clustering [11], and parallel coordinates [12].

In complement to visual-aids, we use table-charts that report the main cluster characteristics and that allow cross-comparison between cluster results. We address the first aspect using the log of the odds which we express for a cluster k on a factor l as $logodds_{kl} = \log((A \times D)/(B \times C))$, see Table 1.

Table 1. For each sum score l, we consider a middle value δ_l such as the data set mean or median. For cells A and B, we use it to count how many observations i in the cluster S_k have a sum score above and below its value. For cells C and D, we proceed to a similar count but on the rest of the observations $i \in \{S - S_k\}$.

	$x_i < \delta_l$	$x_i \geq \delta_l$
$i \in S_k$	A	B
$i \in \{S - S_k\}$	C	D

We address the second aspect using regular association tables. From these tables, the χ^2-statistic is calculated to draw a single association measure in terms of the Cramer's V nominal association coefficient. It expresses as $V = \sqrt{\chi^2/(n \times m)}$ where n is the sample size and $m = min(rows, columns) - 1$. It takes values in $[0, 1]$, one stands for completely correlated variables and zero for stochastically independent ones.

As we perform unsupervised analyses, it is important to know whether the cluster result generalizes to the total patient population. We address this aspect from the machine learning point of view by measuring the classification accuracy of machine learning algorithms like naive Bayes, linear Support Vector Machines or one nearest neighbor as a baseline.

Finally, when conducting a subtype discovery analysis, a key concern is the cluster evaluation. For that purpose, we implemented a simple mechanism to add study-specific evaluation procedures of the subtypes. In OA for instance, as the study involves siblings pairs, we defined two statistical tests that assess the level of familial aggregation in each subtype and its significance.

3 The package, its implementation and a sample analysis

Package design. The implementation articulates around three main containers: the data set `cdata`, the cluster model `cmodel` and the set of cluster results `cresult`. Their entity-relationship cardinalities is as follows: a `cresult` describes a SubtypeDiscovery analysis, it holds a data set `cdata` and it holds several cluster models `cmodel`. In Fig. 2, we illustrate `cdata` requiring an input data set and a description of how it should be interpreted into `settings`. We also describe the relation between `cdata`, `cmodel` and `cresult`.

Plotting a `cdata` container gives for each variable its boxplot, histogram and information like, e.g. its empirical mean or standard deviation. Regards `cresult`, plotting can be restricted to a queried `cmodel` or, by default, it plots all of them. We illustrate such plot for the top-ranking model (VVI, 6, 6022) in Fig. 3. Finally, a print on a `cresult` generates a report that includes the different table charts from the BIC analysis and those focusing on the top-ranking cluster results characterististics, two-by-two comparison, and evaluation. We report some of the most important table-charts in Table 2.

Public `wada2008` data set and sample analysis. Originally generated by Edward O. Cannon, the data set is composed of substances taken from the 2008 WADA (World Anti-Doping Agency) Prohibited List together with molecules having similar biological activity and chemical structure from the MDL Drug Data Report database. Those molecules may belong to ten different activity classes: the β blockers, anabolic agents, hormones and related substances, β-2 agonists, hormone antagonists and modulators, diuretics and other masking agents, stimulants, narcotics, cannabinoids and glucocorticosteroids. This list of molecules was imported into Molecular Operating Environment (MOE) from which all 184 two dimensional descriptors were calculated. The `wada2008` data set is similar to the `wada2005` which was previously published in [13].

```
library(SubtypeDiscovery)
# LOAD DATA SET
data(wada2008)
data(wada2008_settings)
# PREPARE CDATA
cdata1  <- set_cdata(data=wada2008,
   prefix="WADA2008_Sample_Analysis", settings=wada2008_settings)
# PREPARE NEW CDATA FOR CANALYSIS ON PRINCOMP (EXPL. 95\% OF THE VAR.)
cdata2 <- get_cdata_princomp(cdata1)
# PREPARE THE SET OF RESULTS FOR CLUSTER MODELLING
x <- set_cresult(cdata=cdata2, fun_pattern=list(mean=patternMean)
   cfun_settings=list(modelName=c("EII","VII","EEI","VEI","EVI","VVI"),
      G=3:6, rseed=6013:6024))
# PROCEED TO THE MODELLING, SAVING, BIC ANALYSIS, PLOT, PRINT AND WRITE MODELS
x <- analysis(x)
```

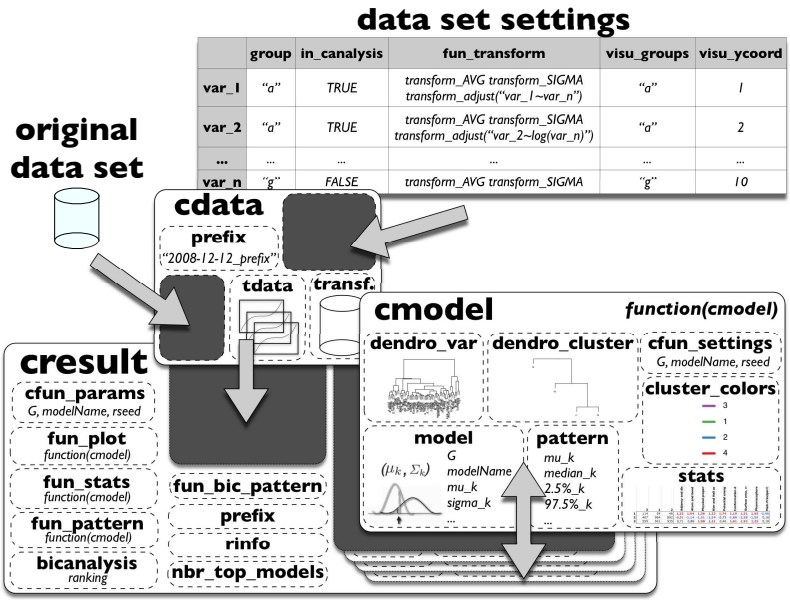

Fig. 2. In the top figure, we illustrate graphically the making of a `cdata` data set container by `set_cdata()` which takes as input raw data and `settings`. These settings describe in particular the sequence of transformation to apply on the data. Next, we report the main three classes of our package, i.e. the data container `cdata`, the model container `cmodel` and the set of cluster results container `cresult`.

4 Concluding remarks

We developed a methodology to facilitate and enhance the search for more homogeneous subtypes with application to medical research and chemoinformatics.

Table 2. Given all repeats, we report tables with the best BIC scores, the starting values leading the most likely model, the ranking of the models given a number of clusters and vice versa and the cross comparison of cluster results (VVI, 6, 6022) and (VVI, 6, 6016). Then, on factor A/d, A/B, C.1, C.2, KHK, Q, Pha., Phys., numbers above 1 and below -1 illustrate high odd ratios especially characterizing a subtype, whereas the association table and the Cramer's V measure (65%) illustrate the level of association between the results.

	EII	**VII**	**EEI**	**VEI**	**EVI**	**VVI**
3	-172857.6	-137398.7	-173010.0	-120257.3	-133262.9	-109035.2
4	-173025.8	-131224.6	-173178.1	-114099.2	-129299.2	-104159.8
5	-173194.1	-127733.5	-173346.4	-109503.1	-126058.9	-99545.5
6	-173362.1	-124697.7	-173514.5	-105724.1	-122780.8	**-93887.8**

	EII	**VII**	**EEI**	**VEI**	**EVI**	**VVI**
3	6013	6013	6013	6022	6017	6014
4	6013	6013	6013	6023	6017	6017
5	6016	6019	6016	6018	6020	6017
6	6016	6022	6016	6016	6024	**6022**

	EII	**VII**	**EEI**	**VEI**	**EVI**	**VVI**
3	**1**	4	**1**	4	4	4
4	2	3	2	3	3	3
5	3	1.9	3	2	2	2
6	4	**1.1**	4	**1**	**1**	**1**

	3	**4**	**5**	**6**
EII	5	5	5	5
VII	3.7	3.8	3.9	4
EEI	6	6	6	6
EVI	3.3	3.3	3.1	3
VEI	1.8	2	2	2
VVI	**1.3**	**1**	**1**	**1**

	1	2	4	6	3	5	**A/d**	**A/B**	**C.1**	**C.2**	**KHK**	**Q**	**Pha.**	**Phy.**
2	824					227	- .4	- .8	.93	Inf	**-1.3**	**-2.6**	**-1.6**	**-1.4**
5	15	13	7	42	494	446	.8	**1.5**	**-1.5**	.5	**1.7**	0.5	**1.5**	**1.4**
1		193				243	- .3	- .9	.8	Inf	- .9	**-3.3**	- .8	.0
3	177					55	**-3.8**	-Inf	Inf	- .5	-Inf	-Inf	-Inf	-Inf
4		5	156	60	27		.9	**1.6**	**-1.3**	**-2.3**	**1.5**	**2.1**	**2.0**	**2.1**
6				53			-Inf	-Inf	Inf	-Inf	-Inf	Inf	-Inf	-Inf
A/d	- .9	- .6	.4	**1.4**	- .0	.6								
A/B	**-1.5**	- .6	**1.2**	**1.1**	**1.4**	.1								
C.1	**1.5**	.7	- .8	**-1.3**	**-1.2**	- .2								
C.2	**1.2**	Inf	- .7	-7.8	Inf	**2.7**								
KHK	**-1.9**	- .9	1.0	**1.4**	**1.5**	.2								
Q	**-3.7**	**-1.5**	.8	**5.5**	- .0	- .7								
Pha.	**-2.2**	- .6	**1.2**	**1.4**	**1.5**	.0								
Phys.	**-2.1**	**-1.1**	**1.8**	**1.1**	**1.4**	.5								
χ^2	0													
V	65%													

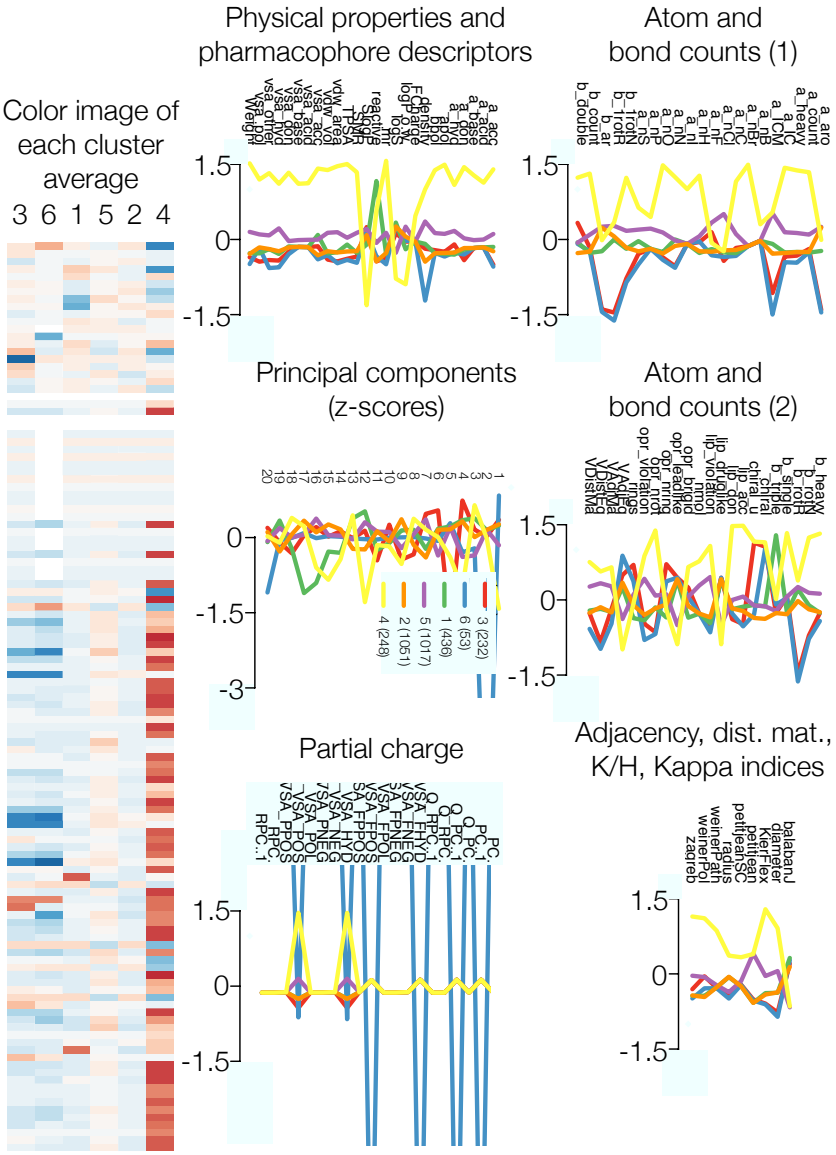

Fig. 3. This Figure exhibits a color image illustrating the six average pattern of (VVI, 6, 6022). It also characteristizes the different subtypes on all variables which we grouped by factor. The plot-scale refers to the z-scores with 95% of the values that should fit within $[-2, 2]$. In this Figure, the yellow subtype with (248) molecules displays an especially high profile on most descriptors. In the contrary, the blue (53) and red (232) subtypes show comparatively low profiles. These two subtypes differentiate on the Partial charge factor where we may account the blue zigzag pattern to the type of the variables which are scores.

In this context, to enable *reproducibility* of our analyses, we release and documented this methodology as the R SubtypeDiscovery package. In this paper, we presented the package implementation and we illustrated its output on an example from chemoinformatics. Ongoing research focuses on the stability of the cluster results given different random starts or when noise is added to data. Parts of the package are also regularly revised or improved aiming for a more reliable and usable methodology.

Acknowledgements This work has been supported by the Netherlands Bioinformatics Centre (NBIC) through its research program BioRange, the Michael J Fox Foundation, PD-subtypes program, while the Leiden University Medical Centre, the Dutch Arthritis Association and Pfizer Inc., Groton, CT, USA support the GARP study (OA).

References

1. Colas, F., Meulenbelt, I., Houwing-Duistermaat, J., van Rooden, S., Visser, M., Marinus, H., van Hilten, B., Slagboom, P.E., Kok, J.N.: Stability of clusters for different time adjustments in complex disease research. In: 30th Annual International IEEE EMBS Conference (EMBC'08), Vancouver, British Columbia, Canada. (August 2008)
2. Colas, F., Meulenbelt, I., Houwing-Duistermaat, J.J., Kloppenburg, M., Watt, I., van Rooden, S.M., Visser, M., Marinus, H., van Hilten, J.J., Slagboom, P.E., Kok, J.N.: A scenario implementation in r for subtype discovery examplified on chemoinformatics data. In: 3rd International Symposium on Leveraging Applications of Formal Methods, Verification and Validation ISOLA, Porto Sani, Greece. (October 2008)
3. Fraley, C., Raftery, A.E.: MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics (September 2006)
4. Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics **49** (1993) 803–821
5. Kass, R.E., Raftery, A.E.: Bayes factors. Journal of the American Statistical Association **90**(430) (1995) 773–??
6. Tukey, J.W.: Exploratory Data Analysis. MA: Addison-Wesley. (1977)
7. Tufte, E.R.: The Visual Display of Quantitative Information. Graphics Press, Cheshire, Connectictut Connectictut (1983)
8. Tufte, E.R.: Envisioning Information. Graphics Press, Cheshire, Connectictut (1990)
9. Brewer, C.A.: 7. In: Color Use Guidelines for Mapping and Visualization. Elsevier Science, Tarrytown, NY (1994) 123–147
10. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. In: Proceedings of National Academy of Science USA. Volume 95. (1998) 11863–14868
11. Sneath, P.H.A., Sokal, R.R.: Numerical Taxonomy, The Principles and Practice of Numerical Classification. Books in Biology. W. H. Freeman and Company (1973)
12. Inselberg, A.: The plane with parallel coordinates. The Visual Computer **1**(2) (1985) 69–91
13. Cannon, E.O., Nigsch, F., Mitchell, J.B.O.: A novel hybrid ultrafast shape descriptor method for use in virtual screening. Chemistry Central Journal **2** (2008)