

Package ‘clustvarel’

November 19, 2015

Type Package

Version 2.2

Date 2015-11-19

Title Variable Selection for Model-Based Clustering

Description A function which implements variable selection methodology for model-based clustering which allows to find the (locally) optimal subset of variables in a data set that have group/cluster information. A greedy or head-long search can be used, either in a forward-backward or backward-forward direction, with or without sub-sampling at the hierarchical clustering stage for starting MCLUST models. By default the algorithm uses a sequential search, but parallelisation is also available.

Author Nema Dean, Adrian E. Raftery, and Luca Scrucca

Maintainer Luca Scrucca <luca@stat.unipg.it>

Depends R (>= 3.2), mclust (>= 5.1), BMA (>= 3.18), foreach, iterators

Suggests MASS, parallel, doParallel

License GPL (>= 2)

Repository CRAN

ByteCompile true

LazyLoad yes

NeedsCompilation no

Date/Publication 2015-11-19 17:21:28

R topics documented:

clustvarel 2

Index 7

Description

A function which implements variable selection methodology for model-based clustering which allows to find the (locally) optimal subset of variables in a dataset that have group/cluster information. A greedy or headlong search can be used, either in a forward-backward or backward-forward direction, with or without sub-sampling at the hierarchical clustering stage for starting Mclust models. By default the algorithm uses a sequential search, but parallelisation is also available.

Usage

```
clustvarsel(data,
            G = 1:9,
            search = c("greedy", "headlong"),
            direction = c("forward", "backward"),
            emModels1 = c("E", "V"),
            emModels2 = mclust.options("emModelNames"),
            samp = FALSE,
            sampsize = round(nrow(data)/2),
            hcModel = "VVV",
            allow.EEE = TRUE,
            forcetwo = TRUE,
            BIC.diff = 0,
            BIC.upper = 0,
            BIC.lower = -10,
            itermax = 100,
            parallel = FALSE)
```

Arguments

data	A numeric matrix or data frame where rows correspond to observations and columns correspond to variables. Categorical variables are not allowed.
G	An integer vector specifying the numbers of mixture components (clusters) for which the BIC is to be calculated. The default is G = 1:9.
search	A character vector indicating whether a "greedy" or, potentially quicker but less optimal, "headlong" algorithm is to be used in the search for clustering variables.
direction	A character vector indicating the type of search: "forward" starts from the empty model and at each step of the algorithm adds/removes a variable until the stopping criterion is satisfied; "backward" starts from the model with all the available variables and at each step of the algorithm removes/adds a variable until the stopping criterion is satisfied.
emModels1	A vector of character strings indicating the models to be fitted in the EM phase of univariate clustering. Possible models are "E" and "V", described in mclustModelNames .

emModels2	A vector of character strings indicating the models to be fitted in the EM phase of multivariate clustering. Possible models are those described in mclustModelNames .
samp	A logical value indicating whether or not a subset of observations is to be used in the hierarchical clustering phase used to get starting values for the EM algorithm.
sampsize	The number of observations to be used in the hierarchical clustering subset. By default, a random sample of approximately half of the sample size is used.
hcModel	A character string specifying the model to be used in hierarchical clustering for choosing the starting values used by the EM algorithm. By default, the "VVV" covariance structure is used (see hc).
allow.EEE	A logical value indicating whether a new clustering will be run with equal within-cluster covariance for hierarchical clustering to get starting values, if the clusterings with variable within-cluster covariance for hierarchical clustering do not produce any viable BIC values.
forcetwo	A logical value indicating whether at least two variables will be forced to be selected initially, regardless of whether BIC evidence suggests bivariate clustering or not.
BIC.diff	A numerical value indicating the minimum BIC difference between clustering and no clustering used to accept the inclusion of a variable in the set of clustering variables in a forward step of the greedy search algorithm. Furthermore, minus BIC.diff is used to accept the exclusion of a selected variable from the set of clustering variable in a backward step of the greedy search algorithm. Default is 0.
BIC.upper	A numerical value indicating the minimum BIC difference between clustering and no clustering used to select a clustering variable in the headlong search. Default is 0.
BIC.lower	A numerical value indicating the level of BIC difference between clustering and no clustering below which a variable will be removed from consideration in the headlong algorithm. Default is -10.
itermax	An integer value giving the maximum number of iterations (of addition and removal steps) the selected algorithm is allowed to run for.
parallel	This argument allows to specify if the selected "greedy" algorithm should be run sequentially or in parallel. The possible values are: <ul style="list-style-type: none">• a logical value specifying if parallel computing should be used (TRUE) or not (FALSE, default) for running the algorithm;• a numerical value which gives the number of cores to employ. By default, this is obtained from the function detectCores;• a character string specifying the type of parallelisation to use. This depends on system OS: on Windows OS only "snow" type functionality is available, while on Unix/Linux/Mac OSX both "snow" and "multicore" (default) functionalities are available.

Clearly, also the last two options imply that the search is performed in parallel.

Details

This function implements variable selection methodology for model-based clustering. The selection methods available are "greedy" search and "headlong" search (see argument `search`). The *greedy search* at each step either checks all the variables not currently included in the set of clustering variables singly for inclusion into the set, or checks all variables in the set of clustering variables singly for exclusion. The *headlong search* only checks until a variable is included or excluded (i.e., does not necessarily check all possible variables for inclusion/exclusion at each step) and any variable with evidence of clustering below a certain level at any stage is removed from consideration for the remainder of the algorithm. Greedy search can be performed *forward-backward* starting from the empty model, or *backward-forward* starting from the model with all the variables included (see argument `direction`). Currently, headlong search can only run *forward-backward*.

The criterion to assess each variable's evidence for being useful to the clustering is given by the difference between the BIC for the clustering model using the set of clustering variables already included and the variable being checked, and the sum of BICs for the model with clustering using the set of already selected clustering variables without the variable being checked and the model for the variable being checked being conditionally independent of the clustering given the other clustering variables. The latter is modeled as a regression of the variable being checked on the other clustering variables. A subset selection procedure is also performed on the regression step.

Clustering models are fitted using `Mclust`, where each model is allowed to vary over the specified number of components `G` and any of the different covariance parameterisations set in `emModels1` and `emModels2`.

The default value for `forcetwo` is `TRUE` because often in practice there will be little evidence of clustering on the univariate or bivariate level although there is multivariate clustering present and these variables are used as starting points to attempt to find this clustering, if necessary being removed later in the algorithm.

The default value for `allow.EEE` is `TRUE` but if necessary to speed up the algorithm it can be set to `FALSE`. Other speeding-up restrictions include reducing the `emModels1` (to "E", say) and the `emModels2` to a smaller set of covariance parameterisations. Reducing the maximum possible number of clusters present in the data will also increase the speed of the algorithm. Another time-saving device is the `samp` option which uses the same algorithm but uses only a subset of the observations in the expensive hierarchical phase for initialisation of EM algorithm in `Mclust`. Finally, the headlong search may be quicker than the greedy search option in data sets with large numbers of variables (depending on the values of the upper and lower bounds chosen for the BIC difference).

The defaults for the `eps`, `tol` and `itmax` options for the `Mclust` steps run in the algorithm can be changed by using the `mclust.options` function.

Value

An object of class 'clustvarel' with the following components:

- | | |
|-------------------------|--|
| <code>variables</code> | The name of input variables. |
| <code>subset</code> | A vector of values specifying for the selected variables the columns position in the input <code>data.frame</code> or matrix. |
| <code>steps.info</code> | A matrix with a row for each step of the algorithm providing: <ul style="list-style-type: none"> - the name of the variable proposed; |

- the BIC of the clustering variables' model at the end of the step;
- the BIC difference between clustering and not clustering for the variable;
- the type of step (Add/Remove);
- the decision for the proposed step (Accepted/Rejected).

search A string specifying the type of search employed.
direction A string specifying the direction of search employed.

Author(s)

N. Dean, A. E. Raftery, and L. Scrucca.

References

- Raftery, A. E. and Dean, N. (2006) Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101(473), 168-178. <http://www.stat.washington.edu/www/research/reports/2004/tr452.pdf>
- Badsberg, J. H. (1992) Model search in contingency tables by CoCo. In Dodge, Y. and Whittaker, J. (Eds.), *Computational Statistics*, Volume 1, pp. 251-256
- Maugis, C., Celeux, G., Martin-Magniette M. (2009) Variable Selection for Clustering With Gaussian Mixture Models. *Biometrics*, 65(3), 701-709.
- Scrucca, L. and Raftery, A. E. (2015) clustvarsel: A Package Implementing Variable Selection for Model-based Clustering in R. Submitted to *Journal of Statistical Software*. Pre-print available at <http://arxiv.org/abs/1411.0606>.

See Also

[Mclust](#)

Examples

```
# Simulate data with 2 clusters in the first two variables and no
# clustering in the rest. Clusters have mixing proportion pro, means
# mu1 and mu2 and variances sigma1 and sigma2
require(MASS)
n = 200
pro = 0.5
mu1 = c(0,0)
mu2 = c(3,3)
sigma1 = matrix(c(1,0.5,0.5,1),2,2,byrow=TRUE)
sigma2 = matrix(c(1.5,-0.7,-0.7,1.5),2,2,byrow=TRUE)
X = matrix(0, n, 5)
colnames(X) = paste("X", 1:ncol(X), sep = "")
# generate the clustering variables
u = runif(n)
Class = ifelse(u < pro, 1, 2)
X[u < pro, 1:2] = mvrnorm(sum(u < pro), mu = mu1, Sigma = sigma1)
X[u >= pro, 1:2] = mvrnorm(sum(u >= pro), mu = mu2, Sigma = sigma2)
# generate the non-clustering variables
```

```
X[,3] = X[,1] + rnorm(n)
X[,4] = rnorm(n, mean = 1.5, sd = 2)
X[,5] = rnorm(n, mean = 2, sd = 1)
# plot the data
clPairs(X, Class, gap = 0)

# sequential forward greedy search (default)
out = clustvarel(X, G = 1:5)
out
# look at the clustering produced by the selected variables
Xs = X[,out$subset]
mod = Mclust(Xs, G = 1:5)
summary(mod)
table(Class, mod$classification)

## Not run:
# sequential backward greedy search
clustvarel(X, G = 1:5, direction = "backward")

# sequential backward greedy search with subsampling at hierarchical
# intialisation stage
clustvarel(X, G = 1:5, direction = "backward",
           samp = TRUE, sampsize = 50)

# parallel backward greedy search
clustvarel(X, G = 1:5, direction = "backward", parallel = TRUE)

# headlong search
clustvarel(X, G = 1:5, search = "headlong")

## End(Not run)
```

Index

*Topic **cluster**

clustvarsel, 2

clustvarsel, 2

detectCores, 3

hc, 3

Mclust, 4, 5

mclust.options, 4

mclustModelNames, 2, 3

print.clustvarsel (clustvarsel), 2