Tools and methods for model-based clustering in R

Bettina Grün

Rennes 2018
Cluster analysis

The task of grouping a set of objects such that

- Objects in the same group are as similar as possible and
- Objects in different groups are as dissimilar as possible.

The aim is to determine a partition of the given set of objects, e.g., to determine which objects belong to the same group and which to different groups.

Statistical methods:

- Heuristic methods: hierarchical clustering, partitioning methods (e.g., $k$-means).
- Model-based methods: finite mixture models.
Specifying the cluster problem

- The cluster problem is in general perceived as ill defined.
- Different notions of what defines a cluster exist:
  - Compactness.
  - Density-based levels.
  - Connectedness.
  - Functional similarity.
- Several cluster solutions might exist for a given data set depending on which notion is used.
- The application context is important to define which clusters should be targeted and to assess the usefulness of a clustering solution.
Specifying the cluster problem / 2
Model-based clustering methods

- Model-based clustering embeds the clustering problem in a probabilistic framework.
- This implies:
  - Statistical inference tools can be used.
  - Different cluster distributions can be used depending on the cluster notion.
  - More explicit specification of what defines a cluster required than for heuristic methods.
Finite mixture models

- Generative model for observations \((y_i, x_i), i = 1, \ldots, n:\)
  1. Draw a cluster membership indicator \(S_i\) from a multinomial distribution with parameters \(\eta = (\eta_1, \ldots, \eta_K)\).
  2. Draw \(y_i\) given \(x_i\) and \(S_i\) from the cluster distribution:

\[
y_i | x_i \sim f_{S_i}(y_i | x_i).
\]

- The distribution of \((y_i, x_i)\) is then given by

\[
y_i | x_i \sim \sum_{k=1}^{K} \eta_k f_k(y_i | x_i),
\]

where

- \(\eta_k \geq 0\) for all \(k\) and \(\sum_{k=1}^{K} \eta_k = 1\).
- \(f_k()\) represents the cluster distribution.
Finite mixture models / 2

Methods differ with respect to:

- **Clustering kernel:**
  - Specification of cluster distributions.
  - Use of additional variables $x_i$, e.g., for regression.

- **Estimation framework:**
  - Maximum likelihood estimation.
  - Bayesian estimation.
Cluster membership indicators can be inferred using the a-posteriori probabilities:

\[ P(S_i = k | y_i, x_i) \propto \eta_k f_k(y_i | x_i). \]

A hard assignment can be obtained by
- Assigning to the cluster where this probability is maximum.
- Drawing from this probability distribution.
Finite mixture models / 4
Estimation of finite mixtures with fixed $K$

- Maximum likelihood estimation:
  - EM algorithm.
  - General purpose optimizers.
  - Hybrid approaches.

- Bayesian estimation:
  - MCMC sampling with data augmentation by adding $S_i$, $i = 1, \ldots, n$.
  - General purpose Gibbs samplers can be used, e.g., JAGS available in R through package rjags (Plummer, 2016).
**EM algorithm**

- Standard maximum likelihood estimation method in a missing data context.
- Guaranteed to converge for bounded likelihoods.
- Only convergence to a local optimum.
- In general slow convergence behavior.
- Consists of E- and M-step:
  - E-step requires determining the a-posteriori probabilities.
  - M-step requires weighted maximum likelihood estimation of the cluster distributions.
MCMC sampling

- Determination of the a-posteriori probabilities required to draw $S_i$, $i = 1, \ldots, n$ from a multinominal distribution.
- Conditional on $S_i$, $i = 1, \ldots, n$ drawing from the posterior of the cluster-specific parameters is the same as if the cluster-specific distribution is used for the complete data set.
- Often poor mixing observed.
- For symmetric priors the posterior is also symmetric and thus multimodal.
Determining the number of clusters

- No generally accepted solution available.
- Suggested methods include:
  - Information criteria: AIC, BIC, ICL.
  - Likelihood ratio test with distribution under the null determined using sampling methods.
  - Marginal likelihoods in Bayesian estimation.
Clustering kernel

- **Components corresponding to clusters:**
  In general using parametric distributions for the components and thus also for the clusters.
  - Multivariate continuous data.
  - Multivariate categorical data.
  - Multivariate mixed data.
  - Multivariate data with regression structure.

- **Combining components to clusters:**
  I.e., the cluster distributions are mixture distributions.
  - Two-step procedures.
  - Simultaneous estimation using constraints or informative priors.

In the following these variants are investigated for maximum likelihood estimation.
Multivariate continuous data

- The standard model is a mixture of multivariate Gaussians.
- The model-based clustering model is given by

\[ y_i \sim \sum_{k=1}^{K} \eta_k \phi(y_i | \mu_k, \Sigma_k). \]

- For \( K \) clusters and \( d \)-dimensional observations \( y_i \) the number of estimated parameters corresponds to

\[ K \cdot (d + d(d + 1)/2) + K - 1. \]
Multivariate continuous data / 2

- Parsimony is achieved based on the decomposition of the variance-covariance matrix into
  - Volume $\lambda$.
  - Orientation $D$.
  - Shape $A$.

  given by

  $$\Sigma_k = \lambda_k D_k A_k D_k^\top.$$ 

- 14 different models emerge by imposing different constraints on the variance-covariance matrices within or across clusters.

- Available packages in R, e.g.:
  - `mclust` (Scrucca et al., 2016),
  - `mixture` (Browne et al., 2018),
  - `Rmixmod` (Lebret et al., 2015).
Multivariate continuous data

[Diagram showing multivariate data with different clusters]
Alternative approaches to achieve parsimony are mixtures of factor analyzers.

- E.g., package `pgmm` (McNicholas et al., 2018) in R.

If the cluster shapes are not symmetric and light tailed, alternative cluster kernels are:

- $t$-distributions (e.g., package `teigen`; Andrews et al. 2018).
- Skewed and / or heavy tailed distributions: e.g.,
  - `EMMIXcskew` (Lee and McLachlan, 2018),
  - `MixSAL` (Franczak et al., 2018),
  - `mixsmsn` (Prates et al., 2013).
Multivariate categorical data

- Often also referred to as latent class analysis.
- Clusters induce a dependency between variables, while variables are independent within clusters.
  ⇒ Local independency assumption.
- The model-based clustering model is given by

\[ y_i \sim \sum_{k=1}^{K} \eta_k \left[ \prod_{j=1}^{d} \text{Multinomial}(y_{ij} | \pi_{jk}^i) \right] \]

for \(d\)-dimensional observations.
- Available packages in R: e.g.,
  - **polCA** (Linzer and Lewis, 2011)
  - **Rmixmod** (Lebret et al., 2015)
Multivariate data with regression structure

- Often also referred to as clusterwise regression.
- The model-based clustering model is given by

\[ y_i | x_i \sim \sum_{k=1}^{K} \eta_k f(y_i | \mu_k(x_i), \phi_k). \]

- Different regression models possible:
  - Generalized linear models.
  - Generalized linear mixed-effects models.
- Available packages in R: e.g.,
  - **flexmix** (Leisch, 2004; Grün and Leisch, 2008)
  - **mixtools** (Benaglia et al., 2009)
Multivariate data with regression structure
Combining components to clusters

- Two-step procedures:
  1. Fit a mixture model as semi-parametric tool for density estimation.
  2. Combine components of the mixture model to form clusters based on some criterion.

Available packages in R, e.g.:
- mclust uses entropy as criterion (Baudry et al., 2010).
- fpc (Hennig, 2018) provides several variants as proposed in Hennig (2010).

- Simultaneous estimation using informative priors in Bayesian estimation can be used in combination with standard estimation methods.
Post-processing tools

- Inference on partitions.
- Inference on cluster-specific parameters:
  In particular for Bayesian estimation the label switching problem needs to be resolved.
- Assigning new observations to clusters:
  Cluster predictions possible.
- Assessing cluster quality.
Assessing cluster quality

- Agreement measures between cluster assignments and true classes available as in a supervised setting:
  - Label-invariant measures:
    - Rand index (corrected for agreement by chance).
    - Jaccard index.
    - Purity.
  - Label-specific measures:
    - Misclassification rate.

- Available packages in R:
  - Package **clue** (Hornik, 2005) provides general infrastructure to assess cluster solutions:
    Function `cl_agreement` provides several methods to assess cluster agreement.
Comparing packages

- Use a classification data set with 3 known classes and three continuous predictors.
- Compare the three packages for fitting mixtures of Gaussians with variance-covariance matrix decomposition.
- Fit all models for $K = 1, \ldots, 9$ and select the best according to the BIC.
- Extract the cluster assignments and assess cluster quality in comparison to the true classification.
Comparing packages: data

```r
> data("diabetes", package = "mclust")
> library("clue")
> class <- as.cl_partition(diabetes$class)
> X <- diabetes[, -1]
> pairs(X, col = diabetes$class, pch = 19)
```
Comparing packages: data / 2
Comparing packages: mclust

```r
> library("mclust")
> mclust.sol <- Mclust(X, G = 1:9)
> class(mclust.sol)

[1] "Mclust"

> mclust.sol[c("G", "modelName")]

$G
[1] 3

$modelName
[1] "VVV"

> c(logLik = logLik(mclust.sol), BIC = BIC(mclust.sol))

     logLik       BIC
-2303.493 4751.311
```
Comparing packages: mclust / 2

Standard visualization:

> plot(mclust.sol, "BIC")
> plot(mclust.sol, "classification")
> plot(mclust.sol, "uncertainty")
> plot(mclust.sol, "density")
Comparing packages: mclust / 3
Comparing packages: mclust / 4

Class agreement:
> cl_agreement(as.cl_hard_partition(mclust.sol), class, 
+    method = "crand")

Cross-agreements using corrected Rand index:
   [,1]
[1,] 0.6640181
Comparing packages: mixture

```r
> set.seed(1234)
> library("mixture")
> mixture.sol <- gpcm(as.matrix(X), G = 1:9)
> class(mixture.sol)

[1] "gpcm"

> (best <- mixture.sol$bicModel[c("G", "covtype")])

$G
[1] 3

$covtype
[1] "VVV"

> mixture.sol$BIC[best$G, best$covtype, c("loglik", "BIC")]
   loglik      BIC
-2303.492 4751.309
```
Comparing packages: mixture

Standard visualization:

```R
> plot(mixture.sol)
```
Comparing packages: mixture / 3

Class agreement:
> mixture.preds <- mixture.sol$map
> cl_agreement(as.cl_partition(mixture.preds), class,
+   method = "crand")

Cross-agreements using corrected Rand index:
   [,1]
[1,] 0.6640181
Comparing packages: Rmixmod

```r
> library("Rmixmod")
> mixmod.sol.1 <- mixmodCluster(
+   data = X, nbCluster = 1,
+   models = mixmodGaussianModel(listModels =
+     c("Gaussian_pk_Lk_Ck", "Gaussian_pk_Lk_Bk",
+       "Gaussian_pk_Lk_I")))
> mixmod.sol.1["bestResult"]@criterionValue

[1] 5136.446

> mixmod.sol <- mixmodCluster(
+   data = X, nbCluster = 2:9,
+   models = mixmodGaussianModel(equal.proportions = FALSE),
+   strategy = mixmodStrategy(nbTryInInit = 20), seed = 10)
> class(mixmod.sol)

[1] "MixmodCluster"
attr(,"package")
[1] "Rmixmod"
```
Comparing packages: Rmixmod

> mixmod.sol["bestResult"]@nbCluster
[1] 3

> mixmod.sol["bestResult"]@model
[1] "Gaussian_pk_Lk_Ck"

> c(logLik = mixmod.sol["bestResult"]@likelihood,
+     BIC = mixmod.sol["bestResult"]@criterionValue)

   logLik     BIC
-2303.493  4751.311

Standard visualization:
> plot(mixmod.sol)
Comparing packages: Rmixmod / 3
Comparing packages: Rmixmod

Some coding also allows to obtain the BIC values for all fitted models ...
Class agreement:

```r
> mixmod.preds <- mixmod.sol["bestResult"]@partition
> cl_agreement(as.cl_partition(mixmod.preds), class,
+   method = "crand")
```

Cross-agreements using corrected Rand index:

```
[,1]
[1,] 0.6640181
```
Comparing packages

- All packages implement the EM algorithm and allow to specify the number of clusters and which variance-covariance structures should be fitted.
- By default all packages use the BIC to select the best fitting model.
- For this simple example all packages arrive at the same solution.
- However, the packages use different initialization strategies.
- The return objects differ as well as the associated methods available to inspect the obtained solutions.
- Class assignments can be extracted and used in combination with package `clue` to assess cluster quality.
Summary

- Model-based clustering is a versatile method for clustering.
- Different variants exist depending on
  - Clustering kernel.
  - Estimation methods.
- A large number of R packages are available where each covers a certain set of mixture models.
- The same functionality is often covered by several packages.
- Packages vary in the range of models covered and what additional functionality is provided for fitted models.
- Common standards and infrastructure are mostly lacking.
- For more information see the CRAN Task View: Cluster Analysis & Finite Mixture Models:
  https://CRAN.R-project.org/view=Cluster
References


References


