

# Package ‘integrativeME’

March 8, 2010

**Type** Package

**Title** integrative mixture of experts

**Version** 1.2

**Date** 2009-09-29

**Depends** mclust, mixOmics, randomForest

**Author** Kim-Anh Le Cao

**Maintainer** Kim-Anh Le Cao <k.lecao@uq.edu.au>

**Description** Mixture of experts models (Jacobs et al., 1991) were introduced to account for nonlinearities and other complexities in the data. It is based on a divide-and-conquer strategy. Mixture of experts are of interest due to their wide applicability and the advantages of fast learning via the expectation-maximization (EM) algorithm. We have extended and implemented mixture of experts to combine categorical clinical factors and continuous microarray data in a binary classification framework to analyze cancer studies. To provide a hybrid signature of clinical factors and gene markers, we propose to apply different gene selection procedures as a first step.

**License** GPL (>= 2)

**Repository** CRAN

**Date/Publication** 2010-03-07 21:23:45

## R topics documented:

breast . . . . .	2
cns . . . . .	3
integrativeME . . . . .	4
integrativeME-internal . . . . .	6
kmeans.init . . . . .	7
kmeansME . . . . .	8
MEfunctions . . . . .	8
prostate . . . . .	9

<b>Index</b>	<b>11</b>
--------------	-----------

---

breast	<i>Breast cancer data, a subset data set from van de Vijver et al. (2002) study</i>
--------	---

---

### Description

Clinical and a subset of gene expression data from the van de Vijver *et al.* (2002) study.

### Usage

```
data(breast)
```

### Format

A list containing the following components:

`type` vector of length 256 indicating the class of the patients (0 = recurrence, 1 = no recurrence).

`cont` data matrix with 256 rows and 500 columns. The gene expression of 500 randomly sampled transcripts (for memory allocation reasons, see details).

`indep` data matrix with 256 rows and 8 columns. The measurements of 8 clinical variables. The discrete data are suitable for a `MElogreg` or `MEindep` model in the mixture of experts methodology.

`loc` data matrix with 256 rows and 7 columns. The measurements of 7 clinical variables. The discrete data are suitable for a `MEloc` model in the mixture of experts methodology.

`loc.ind` indicates the location variable.

### Details

The data set from van de Vijver *et al.* (2002) contains gene expression of tumors from 256 patients who were all treated by modified radical mastectomy or breast-conserving surgery. The authors also included some patients from the Van 't Veer *et al.* (2002) study and the censored patients were removed. The data were preprocessed and filtered to obtain 5,537 genes spotted on Agilent Hu25K microarrays. Eight prognostic factors were available in the clinical data and categorized as indicated by the authors.

- clinical data: include 8 clinical variables
- microarray data: measure the expression of a subset of 500 randomly chosen transcripts.

For the location model, variables 'posnode' and 'chemotherapy' were merged into a single categorical variable (called the location variable).

### Source

See website from the referred article. The original data with 5,537 transcripts can be downloaded as an .RData file from <http://www.math.univ-toulouse.fr/~lecao/package.html>

## References

- van de Vijver, M.J., He, Y.D., van' t Veer, L.J., Dai, H., Hart, A.A.M., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J. and others (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, **347**, 25, 1999–2009.
- Hunt, L. and Jorgensen, M. (1999). Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, **41**, 2, 154–171.

---

cns	<i>Central Nervous System cancer data, a subset data set from Pomeroy et al. (2002) study</i>
-----	---

---

## Description

Clinical and a subset of gene expression data from the Pomeroy *et al.* (2002) study.

## Usage

```
data(breast)
```

## Format

A list containing the following components:

- `type` vector of length 60 indicating the class of the patients (0 = dead, 1 = alive).
- `cont` data matrix with 60 rows and 500 columns. The gene expression of 500 randomly sampled transcripts (for memory allocation reasons, see details).
- `indep` data matrix with 60 rows and 5 columns. The measurements of 5 clinical variables. The discrete data are suitable for a `MElogreg` or `MEindep` model in the mixture of experts methodology.
- `loc` data matrix with 60 rows and 4 columns. The measurements of 4 clinical variables. The discrete data are suitable for a `MEloc` model in the mixture of experts methodology.
- `loc.ind` indicates the location variable.
- `gene.name` gives more information on the genes in the `cont` data set.

## Details

Medulloblastomas are embryonal tumors of the central nervous system (CNS). Pomeroy *et al.* (2002) investigated this malignant brain tumor of childhood as the response of therapy is difficult to predict. The biopsies of 60 patients were obtained before they received any treatment. The expression level of 7,128 genes were available, as well as five clinical variables

- clinical data: include 5 clinical variables
- microarray data: measure the expression of a subset of 500 randomly sampled transcripts.

For the location model, variables 'Age' and 'Cx' were merged into a single categorical variable (called the location variable).

**Source**

See website from the referred article. The original data with 7,128 transcripts can be downloaded as an .RData file from <http://www.math.univ-toulouse.fr/~lecao/package.html>

**References**

Pomeroy, S.L. and Tamayo, P. and Gaasenbeek, M. and Sturla, L.M. and Angelo, M. and McLaughlin, M.E. and Kim, J.Y.H. and Goumnerova, L.C. and Black, P.M. and Lau, C. and others (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 6870, 436–442.

Hunt, L. and Jorgensen, M. (1999). Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, **41**, 2, 154–171.

---

integrativeME	<i>Integrative mixture of experts (ME) to combine clinical variables and microarray data in a binary classification framework.</i>
---------------	--

---

**Description**

We have implemented and further developed mixture of experts (ME) to combine clinical (categorical) data and microarray data. In order to create a hybrid signature, gene selection is performed in a first step with various approaches. The selected genes and the clinical data are then combined with mixture of experts methodology in a classification framework and the evaluation of the performance is performed via k-fold cross-validation.

**Usage**

```
integrativeME(
  data.cat,
  data.cont,
  type,
  select = c('RF', 'student', 'sPLS'),
  method = c('logreg', 'indep', 'loc', 'cont'),
  loc.ind = NULL,
  keepX = 5,
  ng = 2,
  mode.sPLS = NULL,
  fold = 10,
  kcv = 1                                # number of 10-fold cv
)
```

**Arguments**

data.cat	clinical categorical data. The number of clinical factors is usually very small. Data should be sorted according to the type class vector.
data.cont	microarray data. Data should be sorted according to the type class vector.

<code>type</code>	vector indicating the class of each observation or sample. The vector should be coded 0 and 1 and sorted.
<code>select</code>	gene selection method to be applied in the first step. RF = random forests (wrapper method), student = t-test (filter method) and sPLS = sparse PLS, see also <a href="#">spls</a> to select genes according to the clinical factors.
<code>method</code>	variant of ME to use to combine both types of variables, see also <a href="#">MEfunctions</a> .
<code>loc.ind</code>	if <code>method = 'loc'</code> , then the index of the location variable is needed.
<code>keepX</code>	number of genes to select. Should be set to a small value.
<code>ng</code>	number of experts to use in ME, set by default to 2 (i.e. the number of classes)
<code>mode.sPLS</code>	sPLS mode, 'regression' to select genes according to <code>type</code> , 'canonical' to select genes according to the clinical variables.
<code>fold</code>	number of folds in the cross-validation, by default set to 10
<code>kcv</code>	numbers of runs of the cross-validation.

### Details

Clinical variables should all be categorical and should not contain any 0 values.

The samples in `data.cat` and `data.cont` should be ordered according to class vector `type`.

`method = 'cont'` is for continuous data only, whereas 'logreg', 'indep' and 'loc' combine both continuous and categorical variables together.

If `method = 'logreg'`, some warning messages may appear: 'fitted probabilities numerically 0 or 1 occurred' for binomial GLMs, see Venables & Ripley (2002, pp. 197-8).

### Value

<code>mean.error</code>	classification error rate for each cross-validation run.
<code>mat.predcited</code>	predicted class of each observation and each cross-validation run.

### Author(s)

Kim-Anh Le Cao

### References

- Le Cao et al. (2009), submitted.
- Ng, S.K. and McLachlan, G.J. (2008). Expert Networks with Mixed Continuous and Categorical Feature Variables: a Location Modeling Approach. *Machine Learning Research Progress*, ed. Hanna Peters and Mia Vogel, 1–14
- Hunt, L. and Jorgensen, M. (1999). Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, **41**, 2, 154–171.

### See Also

[MEfunctions](#), [kmeans.init](#)

**Examples**

```
## Not run:
data(prostate)
data.cont = prostate$cont
data.cat = prostate$indep
type = prostate$type
type #check that type is sorted
# gene selection with t-test and ME model with a logistic regression in the gatin network:
res = integrativeME(
  data.cat = data.cat,
  data.cont = data.cont,
  type = type,
  select = 'student',
  method = 'logreg',
  keepX = 5,
  ng = 2,
  fold = 10,
  kcv = 1
)

## End(Not run)
```

---

integrativeME-internal

*internal integrativeMe functions*

---

**Description**

Internal functions in integrativeME

**Details**

These functions are not to be called by the user.

**Author(s)**

Kim-Anh Le Cao

---

`kmeans.init`*Initialization of the parameters with K-means*

---

**Description**

Parameters in `integrativeME` are first initialized via K-means clustering algorithm.

**Usage**

```
kmeans.init(data.cont)
```

**Arguments**

`data.cont` gene expression data, where a small number of genes has been selected beforehand. See [integrativeME](#).

**Details**

The number of clusters in K-means is set by default to 2. K-means is used to initialize the parameters before applying the EM algorithm in `integrativeME`.

**Value**

`prop.kmeans` proportions.  
`means.kmeans` estimated means.  
`var.kmeans` estimated variance-covariance matrix.  
`continue` boolean variable, set to FALSE if the number of observations is not sufficient to determine 2 clusters.

**Author(s)**

Kim-Anh Le Cao

**References**

Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics*, **28**, 100-108.

**See Also**

[integrativeME](#), [MEfunctions](#), [kmeans](#).

---

kmeansME	<i>independence gating function for integrative ME</i>
----------	--

---

### Description

More about what it does (maybe more than one line)

---

MEfunctions	<i>Internal gating functions for integrative Mixture of Experts methodology</i>
-------------	---

---

### Description

Different internal gating functions (or models) are proposed within mixture of experts to integrate gene expression data and clinical data in a binary classification framework.

### Usage

```
MEcont(jcross, train, test, n, nv, ng, indclass, data.cont, prop.kmeans, means.kmea
```

```
MEindep(jcross, train, test, n, nq, nv, ng, indclass, data.cat, data.cont, prop.kme
```

```
MElogreg(jcross, train, test, n, nq, nv, ng, indclass, data.cat, data.cont, prop.km
```

```
MEloc(jcross, train, test, n, nq, nv, ng, indclass, data.cat, data.cont, prop.kmea
```

### Arguments

jcross	which cross validation sample
train	training samples
test	test samples
n	number of observations or samples
nq	number of clinical variables
nv	number of genes, the genes should be selected beforehand, see <a href="#">integrativeME</a> .
ng	number of experts, should be set to 2 for a binary classification problem.
indclass	number of samples of class 0.
data.cat	clinical data (categorical).
data.cont	gene expression data.
prop.kmeans	proportions, initialized with k-means, see also <a href="#">kmeans.init</a> .
means.kmeans	means, initialized with k-means, see also <a href="#">kmeans.init</a> .
var.kmeans	variance-covariance matrix, initialized with k-means, see also <a href="#">kmeans.init</a> .
loc.ind	index of the location variable in the case of the MEloc model.

**Details**

Given a training set and a test set, the parameters in `integrativeME` are learnt via the EM algorithm and then tested. All three getting functions are included in the main program `integrativeME`

**Value**

<code>prop</code>	estimated proportions.
<code>w</code>	weighted variable vector in the expert networks function.
<code>loglik</code>	loglikelihood of the model after several iterations.
<code>mat.gum</code>	main output that is used in <code>integrativeME</code> to predict the class label of each tested observation.

**Author(s)**

Kim-Anh Le Cao

**References**

- Le Cao et al. (2009), submitted.
- Ng, S.K. and McLachlan, G.J. (2008). Expert Networks with Mixed Continuous and Categorical Feature Variables: a Location Modeling Approach. *Machine Learning Research Progress*, ed. Hanna Peters and Mia Vogel, 1–14
- Hunt, L. and Jorgensen, M. (1999). Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, **41**, 2, 154–171.

**See Also**

[integrativeME](#), [kmeans.init](#)

---

<code>prostate</code>	<i>Prostate cancer data, a subset data set from Stephenson et al. (2005) study</i>
-----------------------	--

---

**Description**

Clinical and a subset of gene expression data from the Stephenson *et al.* (2005) study.

**Usage**

```
data(prostate)
```

## Format

A list containing the following components:

`type` vector of length 79 indicating the class of the patients (0 = recurrence, 1 = no recurrence).

`cont` data matrix with 79 rows and 500 columns. The gene expression of 500 randomly sampled transcripts (for memory allocation reasons, see details).

`indep` data matrix with 79 rows and 8 columns. The measurements of 8 clinical variables. The discrete data are suitable for a `MElogreg` or `MEindep` model in the mixture of experts methodology.

`loc` data matrix with 79 rows and 7 columns. The measurements of 7 clinical variables. The discrete data are suitable for a `MEloc` model in the mixture of experts methodology.

`loc.ind` indicates the location variable.

## Details

The data set from Stephenson *et al.* (2005) was built from tissue samples obtained from 79 patients all treated by radical prostatectomy. There were 37 samples which were classified as recurrent and 42 as non-recurrent primary prostate tumor. Samples were snap frozen and gene expression analysis was carried out using the Affymetrix U133A human gene array which has 22,283 features. After a prefiltering step, the analyzed data set contained 7,884 features. The clinical data and microarray data were measured on the same set of 79 patients.

- clinical data: include 8 clinical variables
- microarray data: measure the expression of a subset of 500 randomly sampled transcripts.

For the location model, variables 'semi-vesicle invasion' and 'lymph node involvement' were merged into a single categorical variable (called the location variable).

## Source

The data set was obtained upon request to the authors of the study. The original data with 7,884 transcripts can be downloaded as an .RData file from <http://www.math.univ-toulouse.fr/~lecao/package.html>

## References

- Stephenson, A.J., Smith, A., Kattan, M.W., Satagopan, J., Reuter, V.E., Scardino, P.T. and Gerald, W.L. (2005). Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer*, **104**, 2, 290-298.
- Hunt, L. and Jorgensen, M. (1999). Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, **41**, 2, 154-171.

# Index

## \*Topic **datasets**

- breast, 2
- cns, 3
- prostate, 9

breast, 2

cns, 3

counts (*integrativeME-internal*), 6

initmain.indep  
(*integrativeME-internal*), 6

initmain.loc  
(*integrativeME-internal*), 6

initmain.logreg  
(*integrativeME-internal*), 6

integrativeME, 4, 7-9

integrativeME-internal, 6

kmeans, 7

kmeans.init, 5, 7, 8, 9

kmeansME, 8

main.indep  
(*integrativeME-internal*), 6

main.loc  
(*integrativeME-internal*), 6

main.logreg  
(*integrativeME-internal*), 6

MEcont (*MEfunctions*), 8

MEfunctions, 5, 7, 8

MEindep (*MEfunctions*), 8

MEloc (*MEfunctions*), 8

MElogreg (*MEfunctions*), 8

norm.indep  
(*integrativeME-internal*), 6

norm.loc  
(*integrativeME-internal*), 6

norm.logreg  
(*integrativeME-internal*), 6

prostate, 9

spls, 5