Package ‘sureLDA’

October 14, 2022

Type Package

Title A Novel Multi-Disease Automated Phenotyping Method for the EHR

Version 0.1.0-1


URL https://github.com/celehs/sureLDA

BugReports https://github.com/celehs/sureLDA/issues

License GPL-3

Encoding UTF-8

RoxygenNote 7.1.1

Depends R (>= 3.0), Matrix

Imports pROC, glmnet, MAP, Rcpp, foreach, doParallel

LinkingTo Rcpp, RcppArmadillo

Suggests knitr, rmarkdown

VignetteBuilder knitr

LazyData true

NeedsCompilation yes

Author Yuri Ahuja [aut, cre],
Tianxi Cai [aut],
PARSE LTD [aut]

Maintainer Yuri Ahuja <Yuri_Ahuja@hms.harvard.edu>

Repository CRAN

Date/Publication 2020-11-10 10:00:02 UTC

R topics documented:
sureLDA-package .................................................. 2
simdata ....................................................... 2
sureLDA .................................................... 3
**Description**

Surrogate-guided ensemble Latent Dirichlet Allocation (sureLDA) is a label-free multidimensional phenotyping method. It first uses the PheNorm algorithm to initialize probabilities based on two surrogate features for each target disease, and then leverages these probabilities to guide the LDA topic model to generate phenotype-specific topics. Finally, it combines phenotype-feature counts with surrogates via clustering ensemble to yield final phenotype probabilities.

---

**simdata**

*Simulated Dataset*

**Description**

Click [HERE](#) to view details.

**Usage**

simdata

**Format**

An object of class list of length 6.

**Examples**

str(simdata)
surveLDA

---

**Surrogate-guided ensemble Latent Dirichlet Allocation**

### Description

Surrogate-guided ensemble Latent Dirichlet Allocation

### Usage

```r
surveLDA(
  X,
  ICD,
  NLP,
  HU,
  filter,
  prior = "PheNorm",
  weight = "beta",
  nEmpty = 20,
  alpha = 100,
  beta = 100,
  burnin = 50,
  ITER = 150,
  phi = NULL,
  nCores = 1,
  labeled = NULL,
  verbose = FALSE
)
```

### Arguments

- **X**  
  nPatients x nFeatures matrix of EHR feature counts

- **ICD**  
  nPatients x nPhenotypes matrix of main ICD surrogate counts

- **NLP**  
  nPatients x nPhenotypes matrix of main NLP surrogate counts

- **HU**  
  nPatients-dimensional vector containing the healthcare utilization feature

- **filter**  
  nPatients x nPhenotypes binary matrix indicating filter-positives

- **prior**  
  'PheNorm', 'MAP', or nPatients x nPhenotypes matrix of prior probabilities (defaults to PheNorm)

- **weight**  
  'beta', 'uniform', or nPhenotypes x nFeatures matrix of feature weights (defaults to beta)

- **nEmpty**  
  Number of ‘empty’ topics to include in LDA step (defaults to 10)

- **alpha**  
  LDA Dirichlet hyperparameter for patient-topic distribution (defaults to 100)

- **beta**  
  LDA Dirichlet hyperparameter for topic-feature distribution (defaults to 100)

- **burnin**  
  number of burnin Gibbs iterations (defaults to 50)

- **ITER**  
  number of subsequent iterations for inference (defaults to 150)
phi (optional) nPhenotypes x nFeatures pre-trained topic-feature distribution matrix
nCores (optional) Number of parallel cores to use only if phi is provided (defaults to 1)
labeled (optional) nPatients x nPhenotypes matrix of a priori labels (set missing entries to NA)
verbose (optional) indicating whether to output verbose progress updates

Value

scores nPatients x nPhenotypes matrix of weighted patient-phenotype assignment counts from LDA step
probs nPatients x nPhenotypes matrix of patient-phenotype posterior probabilities
ensemble Mean of sureLDA posterior and PheNorm/MAP prior
prior nPatients x nPhenotypes matrix of PheNorm/MAP phenotype probability estimates
phi nPhenotypes x nFeatures topic distribution matrix from LDA step
weights nPhenotypes x nFeatures matrix of topic-feature weights
Index

* datasets
  simdata, 2

* package
  sureLDA-package, 2

simdata, 2
sureLDA, 3
sureLDA-package, 2