Package 'mclustcomp'

August 27, 2018

Type Package

Title Measures for Comparing Clusters

Version 0.3.1

Description Given a set of data points, a clustering is defined as a disjoint partition where each pair of sets in a partition has no overlapping elements. This package provides 25 methods that play a role somewhat similar to distance or metric that measures similarity of two clusterings - or partitions. For a more detailed description, see Meila, M. (2005) <doi:10.1145/1102351.1102424>.

License GPL (>= 3)

Encoding UTF-8

LazyData true

Imports Rcpp, Rdpack

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 6.1.0

RdMacros Rdpack

NeedsCompilation yes

Author Kisung You [aut, cre]

Maintainer Kisung You <kyou@nd.edu>

Repository CRAN

Date/Publication 2018-08-26 22:42:15 UTC

R topics documented:

`mclustcomp-package` *Measures for Comparing Clusterings*

Description

Given a set of data points D, a clustering $C = (C_1, C_2, ..., C_k)$ is a partition where each pair of sets C_i and C_j has no overlapping elements. **mclustcomp** package provides a collection of methods that play a role similar to *distance* or *metric* in that measures similarity of two clusterings (or, partitions) C and C'. For a more detailed description, see Meila, M. (2005) <doi:10.1145/1102351.1102424>.

`mclustcomp` *Measures for Comparing Clusterings*

Description

Given two partitions or clusterings C_1 and C_2, it returns community comparison scores corresponding with a set of designated methods. Note that two label vectors should be of same length having either numeric or factor type. Currently we have 3 classes of methods depending on methodological philosophy behind each. See below for the taxonomy.

Usage

```
mclustcomp(x, y, types = "all", tversky.param = list())
```

Arguments

`x, y`	vectors of clustering labels
`types`	`"all"` for returning scores for every available measure. Either a single score name or a vector of score names can be supplied. See the section for the list of the methods for details.
`tversky.param`	a list of parameters for Tversky index; `alpha` and `beta` for weight parameters, and `sym`, a logical where `FALSE` stands for original method, `TRUE` for a revised variant to symmetrize the score. Default (alpha,beta)=(1,1).

Value

a data frame with columns `types` and corresponding `scores`.

Category 1. Counting Pairs

TYPE	FULL NAME
'adjrand'	Adjusted Rand index.
'chisq'	Chi-Squared Coefficient.
'fmi'	Fowlkes-Mallows index.
'jaccard'	Jaccard index.
'mirkin'	Mirkin Metric, or Equivalence Mismatch Distance.
'overlap'	Overlap Coefficient, or Szymkiewicz-Simpson coefficient.
'pd'	Partition Difference.
'rand'	Rand Index.
'sdc'	Sørensen–Dice Coefficient.
'smc'	Simple Matching Coefficient.
'tanimoto'	Tanimoto index.
'tversky'	Tversky index. Tanimoto Coefficient and Dice's coefficient are special cases with (alpha,beta) = (1,1) and (0.5,
'wallace1'	Wallace Criterion Type 1.
'wallace2'	Wallace Criterion Type 2.

Category 2. Set Overlaps/Matching

TYPE	FULL NAME
'f'	F-Measure.
'mhm'	Meila-Heckerman Measure.
'mmm'	Maximum-Match Measure.
'vdm'	Van Dongen Measure.

Category 3. Information Theory

TYPE	FULL NAME
'jent'	Joint Entropy
'mi'	Mutual Information.
'nmi1'	Normalized Mutual Information by Strehl and Ghosh.
'nmi2'	Normalized Mutual Information by Fred and Jain.
'nmi3'	Normalized Mutual Information by Danon et al.
'nvi'	Normalized Variation of Information.
'vi'	Variation of Information.

References

Strehl A and Ghosh J (2003). "Cluster Ensembles — a Knowledge Reuse Framework for Combining Multiple Partitions." *J. Mach. Learn. Res.*, **3**, pp. 583–617. ISSN 1532-4435, doi: 10.1162/

153244303321897735, http://dx.doi.org/10.1162/153244303321897735.

Meilă M (2007). "Comparing clusterings—an information based distance." *Journal of Multivariate Analysis*, **98**(5), pp. 873–895. ISSN 0047259X, doi: 10.1016/j.jmva.2006.11.013, http://linkinghub.elsevier.com/retrieve/pii/S0047259X06002016.

Meilă M (2003). "Comparing Clusterings by the Variation of Information." In Goos G, Hartmanis J, van Leeuwen J, Schölkopf B and Warmuth MK (eds.), *Learning Theory and Kernel Machines*, volume 2777, pp. 173–187. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-40720-1 978-3-540-45167-9, http://link.springer.com/10.1007/978-3-540-45167-9_14.

Wagner S and Wagner D (2007). "Comparing Clusterings – An Overview." Technical Report 2006-04, Department of Informatics. http://digbib.ubka.uni-karlsruhe.de/volltexte/1000011477.

Albatineh AN, Niewiadomska-Bugaj M and Mihalko D (2006). "On Similarity Indices and Correction for Chance Agreement." *Journal of Classification*, **23**(2), pp. 301–313. ISSN 0176-4268, 1432-1343, doi: 10.1007/s003570060017z, http://link.springer.com/10.1007/s00357-006-0017-z.

Mirkin B (2001). "Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables." *The American Statistician*, **55**(2), pp. 111–120. ISSN 0003-1305, 1537-2731, doi: 10.1198/000313001750358428, http://www.tandfonline.com/doi/abs/10.1198/000313001750358428.

Rand WM (1971). "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association*, **66**(336), pp. 846. ISSN 01621459, doi: 10.2307/2284239, http://www.jstor.org/stable/2284239?origin=crossref.

Kuncheva L and Hadjitodorov S (2004). "Using diversity in cluster ensembles." In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pp. 1214–1219. ISBN 978-0-7803-8567-2, doi: 10.1109/ICSMC.2004.1399790, http://ieeexplore.ieee.org/document/1399790/.

Fowlkes EB and Mallows CL (1983). "A Method for Comparing Two Hierarchical Clusterings." *Journal of the American Statistical Association*, **78**(383), pp. 553–569. ISSN 0162-1459, 1537-274X, doi: 10.1080/01621459.1983.10478008, http://www.tandfonline.com/doi/abs/10.1080/01621459.1983.10478008.

Dongen S (2000). "Performance Criteria for Graph Clustering and Markov Cluster Experiments." CWI (Centre for Mathematics and Computer Science), Amsterdam, The Netherlands, The Netherlands.

Jaccard P (1912). "THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1." *New Phytologist*, **11**(2), pp. 37–50. ISSN 0028-646X, 1469-8137, doi: 10.1111/j.14698137.1912.tb05611.x, http://doi.wiley.com/10.1111/j.1469-8137.1912.tb05611.x.

Li T, Ogihara M and Ma S (2010). "On combining multiple clusterings: an overview and a new perspective." *Applied Intelligence*, **33**(2), pp. 207–219. ISSN 0924-669X, 1573-7497, doi: 10.1007/s1048900901604, http://link.springer.com/10.1007/s10489-009-0160-4.

Larsen B and Aone C (1999). "Fast and effective text mining using linear-time document clustering." In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 16–22. ISBN 978-1-58113-143-7, doi: 10.1145/312129.312186, http://portal.acm.org/citation.cfm?doid=312129.312186.

Meilă M and Heckerman D (2001). "An Experimental Comparison of Model-Based Clustering Methods." *Machine Learning*, **42**(1), pp. 9–29. ISSN 1573-0565, doi: 10.1023/A:1007648401407, https://doi.org/10.1023/A:1007648401407.

Cover TM and Thomas JA (2006). *Elements of information theory*, 2nd ed edition. Wiley-Interscience, Hoboken, N.J. ISBN 978-0-471-24195-9, OCLC: ocm59879802.

Ana L and Jain A (2003). "Robust data clustering." In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pp. II–128–II–133. ISBN 978-0-7695-1900-5, doi: 10.1109/CVPR.2003.1211462, http://ieeexplore.ieee.org/document/1211462/.

Wallace DL (1983). "Comment." *Journal of the American Statistical Association*, **78**(383), pp. 569–576. ISSN 0162-1459, 1537-274X, doi: 10.1080/01621459.1983.10478009, http://www.tandfonline.com/doi/abs/10.1080/01621459.1983.10478009.

Simpson GG (1943). "Mammals and the nature of continents." *American Journal of Science*, **241**, pp. 1–31. doi: 10.2475/ajs.241.1.1.

Dice LR (1945). "Measures of the Amount of Ecologic Association Between Species." *Ecology*, **26**(3), pp. 297–302. ISSN 00129658, doi: 10.2307/1932409, http://doi.wiley.com/10.2307/1932409.

Segaran T (2007). *Programming collective intelligence: building smart web 2.0 applications*, 1st ed edition. O'Reilly, Beijing ; Sebastapol [CA]. ISBN 978-0-596-52932-1, OCLC: ocn166886837.

Tversky A (1977). "Features of similarity." *Psychological Review*, **84**(4), pp. 327–352. ISSN 0033-295X, doi: 10.1037/0033295X.84.4.327, http://content.apa.org/journals/rev/84/4/327.

Danon L, Díaz-Guilera A, Duch J and Arenas A (2005). "Comparing community structure identification." *Journal of Statistical Mechanics: Theory and Experiment*, **2005**(09), pp. P09008–P09008. ISSN 1742-5468, doi: 10.1088/17425468/2005/09/P09008, http://stacks.iop.org/1742-5468/2005/i=09/a=P09008?key=crossref.5420f964e99dd130e25dd14c3f1af547.

Lancichinetti A, Fortunato S and Kertész J (2009). "Detecting the overlapping and hierarchical community structure in complex networks." *New Journal of Physics*, **11**(3), pp. 033015. ISSN 1367-2630, doi: 10.1088/13672630/11/3/033015, http://stacks.iop.org/1367-2630/11/i=3/a=033015?key=crossref.10a0c9c4b54720787488289cc0fb9f78.

Examples

```
## example 1. compare two identical clusterings
x = sample(1:5,20,replace=TRUE) # label from 1 to 5, 10 elements
y = x                           # set two labels x and y equal
mclustcomp(x,y)                 # show all results

## example 2. selection of a few methods
z = sample(1:4,20,replace=TRUE)           # generate a non-trivial clustering
cmethods = c("jaccard","tanimoto","rand") # select 3 methods
mclustcomp(x,z,types=cmethods)            # test with the selected scores

## example 3. tversky.param
tparam = list()                           # create an empty list
tparam$alpha = 2
tparam$beta  = 3
tparam$sym   = TRUE
mclustcomp(x,z,types="tversky")           # default set as Tanimoto case.
mclustcomp(x,z,types="tversky",tversky.param=tparam)
```

Index