This is code for biclustering method named “RelDenClu: A Relative Density based Biclustering Method for identifying non-linear feature relations

**RelDenClu.m:**

The main method is RelDenClu(d, ObsInMinBase, maxSingleLinkageFeat)

Input:

First input d is normalized data to be biclustered

ObsInMinBase is minimum number of observations in a bicluster

maxSingleLinkageFeat is maximum number of feature sets found for each observation set. Default is M/2, where M is number of features in the data

Output:

Matrices Clubase and Cludim

Clubase Each row is the set of observations in a bicluster

Cludim Each row is the set of feature in a bicluster

Demo.m

Shows how to use RelDenClu

preproFast.m

Is called by RelDenClu(). Finds high density connected cells. Input is data d

Output is ixmat connected regions for each feature pair

bicRMfunc.m

Is called by RelDenClu(). Finds intersection of observations forming dense regions for 3 features

Input is ixmat and ObsInMinBase

Output is impbase and impdim

Getbiclus.m

Is called by RelDenClu(). Finds final biclusters

Input data d, impbase from bicRMfunc, user input maxSingleLinkageFeat

Output

Clubase Each row is the set of observations in a bicluster

Cludim Each row is the set of feature in a bicluster

gridMICuneqmainstatEnt.m

Finds pairwise dependence between features for given data, called by getbiclus

Input data d

Power parameter to decide bin length should be close to 0

Output Triangular matrix with dependence values

Norm1.m

Input data d

Output normalized data d

Norm2.m

Useful for Gaussian data

Input data d

Output normalized data d

Following files are Matlab scripts and data meant to demonstrate RelDenClu.

Thus includes an application to COVID dataset to identify World Development Indicators that affect COVID spread using the data at nascent stage of disease (Data from 31-Jan-2020). The factors were later found to have high correlation with 31-Dec-2020 COVID rates. Related files are as follows.

COVID EXPERIMENT FILES

Two data files that were taken from open data sources.

wdiDat.csv

from

*The World Bank. World development indicators (2015).*

The data provided here is obtained by dropping few columns from original dataset.

owid-covid-dataDec.csv

from

Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track covid-19 in eal time. The Lancet Infect. Dis. 20, 533–534 (2020).

Or

Edouard Mathieu, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Saloni Dattani, Diana Beltekian, Esteban Ortiz-Ospina and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: <https://ourworldindata.org/coronavirus>

The code files include scripts

covidJanDec2020Fast.m

The script reads the data (Covid rates for Jan and WDI features) and finds features relating to spread of covid. It calculates correlation coefficients both Spearman and Pearson for WDI features with January and December covid rates. The following scripts are used by this script.

covRateDec.m

For reading covid data from owid-covid-dataDec.csv

besctaccCoc.m

For finding bicluster with best match.

Accutest.m

For finding accuracy with best bicluster

CREDIT CARD DEFAULTER EXPERIMENT FILES

ccdef.csv

from

Dua, D. & Graff, C. UCI machine learning repository (2017).

<https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

The code file is

testCredBi.m: It reads the data, calculates bicluster from data except label and uses them as additional feature. The naive bayes classification accuracy for original and augmented datasets are calculated using 10-fold cross validation.

DATA GENERATION FOR PROPERTIES OF BICLUSTERING

Random biclustering data can be generated using gendata4prop.m which uses scripts datagen2.m and datagenG.m