Directional association inference challenged by severe dropout in single-cell RNA-seq data CODE

EliĹ ˇka DvoĹ™Ă ˇkovĂ ˇ

README

This README serves as overview or documentation of R code created for Directional association inference challenged by severe dropout in single-cell RNA-seq data project in Bioinformatiocs course 2019.

File Overview

1. methods.R
2. data_generation.R
3. experiments.R

methods.R

This file includes all methods we used in our experiments for inference detection.

run_Chisq

function for Pearson's Chi-squared test. Takes 4 argument:

- tbl - contigency table to be checked for functional dependency.
- output_file - file to save the result in
- g1 - name of the first variabl, set by defaul to "X"
- g2 - name of the first variabl, set by defaul to "Y"

run_corr_test

function for Pearson's correlation test. Takes 4 argument:

- tbl - contigency table to be checked for functional dependency.
- output_file - file to save the result in
- g1 - name of the first variabl, set by defaul to "X"
- g2 - name of the first variabl, set by defaul to "Y"

run_muti

function for Mutual Information test. Takes 4 argument:

- tbl - contigency table to be checked for functional dependency.
- output_file - file to save the result in
- g1 - name of the first variabl, set by defaul to "X"
- g2 - name of the first variabl, set by defaul to "Y"

run_contropy

function for Correlation entropy test. Takes 4 argument:

- tbl - contigency table to be checked for functional dependency.
- output_file - file to save the result in
- g1 - name of the first variabl, set by defaul to “X”
- g2 - name of the first variabl, set by defaul to “Y”

run_FunChisq

function for Functional Chi-squared test. Takes 4 argument:

- tbl - contigency table to be checked for functional dependency.
- output_file - file to save the result in
- g1 - name of the first variabl, set by defaul to “X”
- g2 - name of the first variabl, set by defaul to “Y”

run_all

Runs all the functions in methods.R file mentioned above. Takes 6 arguments:

- tbl - contigency table to be checked for functional dependency.
- chisqr - file to save the result of Pearson’s Chi-squared test, default = “Chi.txt”
- cor - file to save the result of Pearson’s correlation test, default = “cor.txt”
- muti - file to save the result of Mutual Information test, default = “muti.txt”
- con_ent - file to save the result of Conditional Entropy test, default = “contropy.txt”
- funchisq - file to save the result of FUnctional Chi-squared test, default = “Funchisq.txt”

data_generation.R

This file contains all code ralated to the data generation like: simulation of contingency tables,dropout simulation or data diretization.

gen_simulated_dataset

This function generate simulated contingency tables with predefined parameters take as arguments:

- n - number of samples set to 1000 by default
- edges - total number of contingency tables se to 200 by default
- noise - noise of generated tables set to 0.0 by default
- directional - boolean defining the type of generated tables if TRUE (default) half of tables are many-to-one typer and the other half is one-to-many. If FALSE falf of tables are functional and the rest of it is independent
- nrow - dimensiona on the tables - created tables will be always square, default value: 3

dropout

Simulated dropout in a vector of classes defined by capital letters. Arguments:

- x - vector to be transformed
- d - dropout rate, default value = 0.0
- n - number of samples, default value = 1000

gen_data_dropout

Simulates dataset in all contingency tables given in data.frame as first agument. Arguments:

- data - data.frame containg all contingency tables to be modified
- n - number of samles of all contingencz tables, deafult value = 1000
- d - dropout rate to simulate, default value = 0.2

data_discr

Discretizes the given dataset. Arguments:

dataset - dataset to be discretized. * size - size of the dataset to be discretize. -1(default) if all dataset should be processed * dim - dimenzion of data to discretize rows or columns, default value = 2 - columns.

create_table

Creates Person → gene contingency tables from given dataset and returns a data.frame containing them. Arguments:

- dataset - dataset to create the contingecy tables for Person → gene dependency

gen_real_dataset

Loads the real dataset and its gound truth in correct form to be used in our experiments. Arguments:

- sizeP - number of sumples to used for each person. -1 (default) indicates to test all samples for each person
- sizeG - number of genes to used. -1 (default) indicates to test all genes.
- expressed - False if we test directionality, True if we test iferance detection. default value: TRUE.

experiments.R

Code in this file encapsulate all designed datasets,

run_experiment

This function proccess all results by method we used and creates ROC and PR curves. Arguments:

- dataset - data used in experiment
- edges_gt - ground truth list of 1 and 0.
- input_files - array of pths to files containing the results of tested methods.
- names - array of string containing names of methods in the same order as the given files
- title - title of the ROC and PR graphs, default = 0.5

process_real_data

Runs the experiment proccesing the real dataset, both configurations of experiments. Takes one argument:

- pdf_name - file name to save the graphs.

process_simulated_data

Created datatsets by given parameters and procceses both configurations of experiments. Arguments:

- dropout_rate - array of decimal numbersfrom 0 to 1 defining all dropout rates to generate, default value: c(0.2, 0.8, 0.9, 0.99)
- noise_levels - array of decimal numbers from 0 to 1 to defined all noise levels to be aplied to datasets = c(0, 0.1, 0.2, 0.5, 1)

test_sample_size

Run an Experiment of testing the influens of sample size on the methods' performance. Argument:

- samples - an array of integers to defined what sample sizes to test.