README document for manuscript “Bayesian Tree-Structured Two-Level Clustering for Nested Data Analysis”

1. Data

1.1. Abstract

The CIFAR-100 data set are collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. This data set has 100 classes (grouped into 20 superclasses) and each class contains 500 training images and 100 test images.

The scRNA-seq data are collected from nine prostate cancer patients for three types of tissue fractions: "tumor," "involved," and "distal." The data set is available in Gene Expression Omnibus platform with code GSE143791.

1.2. Availability

The CIFAR-100 are publicly available for download via the link https://www.cs.toronto.edu/~kriz/cifar.html or keras module in Python.

The scRNA-seq data are publicly available for download via the online data portal at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143791. No registration is required.

1.3. Description

CIFAR-100 data

Link to data: https://www.cs.toronto.edu/~kriz/cifar.html

Data's contributor(s): Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton

Citation: Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

File format: python version, Matlab version, binary version

In the data preprocessing procedure, we focused on nine classes—chair, chimpanzee, dolphin, lion, maple, oak, oranges, flatfish, and shark—and each class has 500 training images. We used UMAP to reduce the dimension of images to 10. Subsequently, we assigned these 4,500 images to 30 sources with each source having 150 images, and the 30 sources further form three groups with each group having 10 sources, where group 1 contains three classes chair, chimpanzee, and dolphin, group 2 contains three classes lion, maple, and oak, and group 3 contains three classes oranges, flatfish, and shark. The processed data is saved in "rawData.csv".

scRNA-seq data
Link to data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143791

Data's contributor(s): Ninib Baryawno

Citation: Youmna Kfoury, Ninib Baryawno, Nicolas Severe, Shenglin Mei, Karin Gustafsson, Taghreed Hirz, Thomas Brouse, Elizabeth W Scadden, Anna A Igolkina, Konstantinos Kokkaliaris, and as part of the Boston Bone Metastases. Human prostate cancer bone metastases have an actionable immunosuppressive microenvironment. *Cancer Cell*, 2021.

File format: csv files

In the data preprocessing procedure, we focused on the scRNA-seq counts data. We removed genes with zero proportions greater than 95% across cells and external RNA controls consortium (ERCC) spike-in molecules. We then normalized the raw count data via logarithmic transformation accounting for library sizes. Next, we computed the expression variance for each gene, selected the top 100 highly expressed genes, and applied UMAP to reduce the data dimensionality to two.

Version information: we used the data version on the data contributors' last update date Sep 05, 2021.

2. Code

2.1. Abstract

All of the data preprocessing and analysis in this paper were completed using R and Python. The code is provided to conduct preprocessing on the raw data, implement TreeTC via Markov chain Monte Carlo methods, compare against baseline methods, and generate descriptive plots.

2.2. Description

All of the R and Python scripts are available as the supplementary code.
R license information: MIT.
Python license information: PSF.

For R and R packages, we use R version 4.0.2 (2020-06-22, “Taking Off Again”). The used R packages are: (System_preparation.R)

- aricode, version 1.0.0 (https://CRAN.R-project.org/package= aricode)
- clevr, version 0.1.1 (https://CRAN.R-project.org/package= clevr)
- cluster, version 2.1.1 (https://CRAN.R-project.org/package= cluster)
- dplyr, version 1.0.7 (https://CRAN.R-project.org/package=dplyr)
- ggplot2, version 3.3.5 (https://CRAN.R-project.org/package= ggplot2)
- ggvis, version 0.4.7 (https://CRAN.R-project.org/package=ggvis)
- gplots, version 3.1.1 (https://CRAN.R-project.org/package= gplots)
- igraph, version 1.2.6 (https://CRAN.R-project.org/package= igraph)
- mcclust, version 1.0 (https://CRAN.R-project.org/package=mcclust)
- mclust, version 5.4.9 (https://CRAN.R-project.org/package=mclust)
- RColorBrewer, version 1.1-2 (https://CRAN.R-project.org/package=RColorBrewer)
- R6, version 2.5.0 (https://CRAN.R-project.org/package=R6)
- tibble, version 3.1.2 (https://CRAN.R-project.org/package=tibble)
- tidyr, version 1.1.3 (https://CRAN.R-project.org/package=tidyr)
- umap, version 0.2.7.0 (https://CRAN.R-project.org/package=umap)

For Python and Python packages, we use Python version 3.6.11 (27 June 2020). The used Python packages are: (execute the code "pip install -r requirements.txt" in the command)

- numpy, version 1.19.1
- pandas, version 1.1.2
- tensorflow, version 1.14.0
- keras, version 2.3.1
- matplotlib, version 3.3.1
- scikit-learn, version 0.23.2
- PIL, version 7.2.0
- re, version 2.2.1

A MacBook Pro was used for the simulation study in this paper. The details of the computer are:

- Operating system: MacOS Catalina 10.15.5
- CPU: Intel Core i5 2GHz
- RAM: 16GB

The computing platform was used for the real application analyses in this paper. The details of the computing platform are:

- Operating system: CentOS 7.8.2003
- CPU: Intel Gold 5218 (16 cores, 32 threads) 2.3GHz
- RAM: 192GB

2.3. Instructions for Use

All data preprocessing and analysis as well as Figures 4-7 in the manuscript can be reproduced.

Detailed workflow information is contained in the "README.pdf" in "Simulations" and "Real Application" directories. One should firstly check and install the R and Python by conducting the code file "System_preparation.R".

The general steps in the nonoverlapping and overlapping simulation cases are:

1. Generate the data.
2. Apply the TreeTC model to the data and select the iteration following the big node number rule. (There are 30,000 iterations in the MCMC with 5,000 burn-in steps.)
3. Conduct the CAM model, Gaussian mixture model, and two baseline methods.
4. Generate figure 4 in the paper.

The general steps in the first real application (image data) are:

1. Conduct data preprocessing.
2. Apply the TreeTC model to the preprocessed data and select the iteration following the big node number rule. (There are 10,000 iterations in the MCMC with 5,000 burn-in steps.)
3. Generate figure 5 in the paper.

The general steps in the second real application (prostate cancer single-cell RNA-seq data) are:

1. Conduct data preprocessing.
2. Apply the TreeTC model to the preprocessed data and select the iteration following the big node number rule. (There are 10,000 iterations in the MCMC with 5,000 burn-in steps.)
3. Generate figure 6 in the paper.
4. Generate figure 7 in the paper.