Package 'HTT'

August 23, 2024

Type Package

Title Data, scripts, and functions of the High-Throughput Truthing project

Version 2.0.1

Author Brandon D. Gallas

Maintainer <brandon.gallas@fda.hhs.gov>

Description This package contains data collected for the High-Throughput Truthing project (HTT). There are also functions and scripts replicating analyses done for different presentations and publications (in the ``inst" directory).

Imports iMRMC, stats

Depends R (>= 2.10)

License CC0

Encoding UTF-8

LazyData true

RoxygenNote 7.3.1

R topics documented:

`agreeDensityBRBM` *Create and analyze paired-reader data*

Description

Create and analyze paired-reader data

Usage

```
agreeDensityBRBM(
  mrmcDF,
  modalityLabel = "5: all panel",
  subgroupLabel = "1: all considered"
)
```

Arguments

`mrmcDF`	(data.frame) - Data frame for analysis
`modalityLabel`	(chr) - Platform used by the readers. Possible values:("1: camic", "2: eedap", "3: pathp", "4: all", or "5: all panel")
`subgroupLabel`	(chr) - Range of scores analyzed by readers. Possible Values: ("1: all considered", "2: avg. less than or equal to 10", "3: avg. greater than 10 and less than or equal to 40", "4: avg. greater than 40", "5: score.X equal to zero")

Details

This function compares and analyzes the paired scores of all pairs of readers for every annotated region of interest (`caseID`).

The input data frame is filtered according to `modalityLabel`. Then, the function pairs the "between-reader data" and filters it according to `subgroupLabel`. Then for each pair of readers, the function calculates summary statistics, including the average squared differences of the `scores`.

Value

A large list that contains the following elements:

- `modalityLabel` (chr) - Platform used by the readers
- `subgroupLabel` (chr) - Range of scores analyzed by readers
- `modalityID` (chr) - Platform used to analyze slides
- `nR` (int) - Number of total readers
- `readers` (chr) - readerID of all readers
- `xlim.filter` (num) - Range of scores (1-100) analyzed by readers
- `mrmcDF` (data frame) - Data frame that contains the annotation data filtered according to `modalityLabel`. This will be `pilotHTT` or some filtered version
- `mrmcBRWM` (data frame) - Data frame that compares the scores of all paired readers for the same annotated region of interest (caseID). Data contains the following variables:
  1. `caseID` (fact) - ID for region of interest
  2. `readerID.X` (fact) - readerID of reader X

3. modalityID.X (fact) - modality used by reader X
4. score.X (num) - score by reader X
5. readerID.Y (fact) - readerID of reader Y
6. modalityID.Y (fact) - modality used by reader Y
7. score.Y (num) - score by reader Y

- resultsBRWM (data.frame) - Data containing the average squared difference for all pairs of readers. Data contains the following variables:
  1. modalityID (chr) - modality used used to annotate ROI
  2. readerID.1 (chr) - readerID of reader 1
  3. readerID.2 (chr) - readerID of reader 2
  4. nObs (num) - number of paired observations
  5. MSD (num) - Average of the scores' squared differences

Examples

```
results <- agreeDensityBRBM(HTT::pilotHTT)
str(results)
```

---

casesHTT *Original image file names and related information for HTT cases*

Description

This file contains the original image file names and related information for the HTT cases. The data collected from caMicroscope, eeDAP, and pathPresenter.

Usage

```
casesHTT
```

Details

View the image and scanner information of casesHTT: scannerInformationCasesHTT.

Value

The dataframe contains 9 columns:

- batch (chr) - Batch number of image annotated by the reader (8 batches in total)
- scanYear (int) - Year when the slide was scanned
- WSIoriginal (chr) - New whole case file name of whole slide image annotated by reader
- WSInew (chr) - Original whole case file name of whole slide image annotated by reader
- cancerType (chr) - Type of cancer of slides
- sampleType (chr) - Type of tissue sample (biopsy or resection)
- glassSlideRecieved (logical) - If the glass slide of the sample was received by the FDA
- note (logical) - Additional notes
- received (chr) - Method by which images were received

cleanReaders *Information about the readers in this study*

Description

This file contains the information about the readers in this study. It has been cleaned of PII (names and emails)

Usage

```
cleanReaders
```

Details

readerID updated from reader#### to profession### depending on the number of years of experience and experienceResident

This data is saved as rda and csv files

Value

A data fame with the following (3) variables:

- `readerID` (factor) - ID of participant (profession and ID number)
- `experience` (num) - Number of years of experience for pathologists
- `experienceResident` (num) - Number of years in residency for non-pathologists

doStatsByCase *Summarize the scores for each case of a data frame*

Description

This function creates a one row data frame that summarizes the scores for each case of an imported data frame (`mrmcByCase.cur`).

Usage

```
doStatsByCase(mrmcByCase.cur)
```

Arguments

`mrmcByCase.cur`	A data frame with the following variables: batch, WSI, caseID, modalityID, score, labelROI.

Value

A data frame (one row) with the following (11) variables:

- `batch` (factor) - Batch number of image annotated by the reader (8 batches in total)
- `WSI` (factor) - Whole case file name of whole slide image annotated by reader
- `caseID` (factor) - ID for region of interest. Includes WSI, x position of ROI, y position of ROI, and length of ROI
- `modalityID` (factor) - Platform used by viewer (caMicro, pathPresenter, or eeDAP)
- `nObs` (int) - Total number of observations
- `nObs.na` (int) - Number of observations labeled as not evaluable ... there is no score (NA)
- `nObs.evaluable` (int) - Number of observations labeled as evaluable ... a score has been provided
- `scoreMean` (num) - Average percent of area occupied by lymphocytes in tumor-associated stroma
- `scoreVar` (num) - Variance of the percentages of area occupied by lymphocytes in tumor-associated stroma
- `CV` (num) - Coefficient of variation of the percent area occupied by lymphocytes in tumor-associated stroma (sqrt(scoreVar) / scoreMean)
- `labelMajority` (chr) - The majority label of ROI
- `labelEntropy` (num) - The entropy of the observed label distribution

Examples

```
# Get Data
df1 <- HTT::pilotHTT

# Select data from a single caseID (ROI) and a single modalityID
df2 <- df1[df1$caseID == df1$caseID[2] & df1$modalityID == df1$modalityID[1], ]

# Run the function
result <- HTT::doStatsByCase(df2)

# View the result
print(result)
```

`getBlandAltmanWithDuplicates` *Compare the data from two readers (count duplicates).*

Description

This function compares the scores from two readers. In a sense it treats the comparisons as factors and counts the number of times specific pairs of scores are repeated. This accounting allows a Bland-Altman plot to be created with symbols scaled to the number of times each pair of score is repeated.

Usage

```
getBlandAltmanWithDuplicates(mrmcBRWM)
```

Arguments

mrmcBRWM	A data frame with the following (7) variables: • caseID (factor) - ID for region of interest. Includes WSI, x position of ROI, y position of ROI, and length of ROI • modalityID.X (factor) - Platform used by reader X (caMicro, pathPresenter, or eeDAP) • readerID.X (factor) - reader ID of reader X • score.X (num) - scores of reader X • readerID.Y (factor) - reader ID of reader Y • modalityID.Y (factor) - Platform used by reader Y (caMicro, pathPresenter, or eeDAP) • score.Y (num) - scores of reader Y

Value

A data frame with the following (5) variables:

- nobs (int) - A vector of the number of observations
- x (num) - scores of reader x
- y (num) - scores of reader y
- xyAvg (num) - Averages of the scores between the two readers
- xyDiff (num) - Differences of the scores between the two readers

Examples

```
# Created between-reader, between modality paired data
mrmcBRWM <- iMRMC::getBRBM(pilotHTT, "camic", "camic")

# Determine the frequency table for paired observations
resultBlaAlt <- HTT::getBlandAltmanWithDuplicates(mrmcBRWM)

# Show the results
head(resultBlaAlt)
plot(resultBlaAlt$x, resultBlaAlt$y)
```

---

lineFromTwoPoints *Find the Y value at* inputX *of a line from two points*

Description

This function finds the y value at inputX on a line defined by (x1, y1) and (x2, y2)

Usage

```
lineFromTwoPoints(inputX, x1, y1, x2, y2)
```

Arguments

inputX	(int) - The point at which to evaluate the line
x1	(num) - x-value for first point
y1	(num) - y-value for first point
x2	(num) - x-value for second point
y2	(num) - y-value for second point

Value

The y value of the line at `inputX` (num)

Examples

```
x <- 1:100
y <- lineFromTwoPoints(x, 1, 1, 100, 5)
```

---

pilotHTT *Annotation data*

Description

This file is the aggregate of all clean data from the High-Throughput Truthing project. It has been cleaned of PII (names and emails) and other non-essential columns.

Usage

```
pilotHTT
```

Details

This data was collected from the CAmicroscope, PathPresenter, and eeDAP platforms. Please refer to for more information about the data.

As of 6 May 2022, this data contains 7898 observations of 18 variables.

Value

A data frame with the following (18) variables:

- `batch` (factor) - Batch number of image annotated by the reader (10 batches in total)
  - `FDA-HTT-batch00x` - Pilot Study annotations
  - `FDA-HTT-Train00x` - Expert Panel annotations
- `WSI` (factor) - Whole case file name of whole slide image annotated by reader
- `caseID` (factor) - ID for region of interest. Includes WSI, x position of ROI, y position of ROI, and length of ROI
- `readerID` (factor) - ID for each participant (profession with ID number)
  - There are four possible professions at the front end of readerID: pathologist, expert, resident, or unknown.
  - `pathologist` - board-certified pathologist

  - expert - member of the Expert Panel
  - resident - in residency
  - unknown - no indicated profession
- modalityID (factor) - Platform used by viewer (caMicro, pathPresenter, eeDAP, or camic-expert)
  - caMicro - Digital annotations collected
  - pathPresenter - Digital annotations collected
  - eeDAP - Microscope annotations collected
  - camic-expert - Digital Expert Panel annotations collected using the caMicroscope platform
- score (num) - Percent of area occupied by lymphocytes in Intra-Tumoral Stroma. (Same as densityTILs).
- experience (num) - Number of years of experience for pathologists. If experience == 100, experience is unknown
- experienceResident (num) - Number of years in residency for non-pathologists. If experienceResident == 100, experience is unknown
- labelROI (factor) - Label of region of interest (Intra-Tumoral Stroma, Invasive Margin, Tumor with No Intervening Stroma, other regions)
- VTA (logical) - Indicates whether the region of interest is appropriate for sTIL evaluation
- percentStroma (num) - Percentage of tumor-associated stroma in region of interest
- densityTILs (num) - Percent of area occupied by lymphocytes in Intra-Tumoral Stroma. (Same as score)
- createDate (POSIXct) - Date and time annotation was created
- viewerWidth (num) - Width of image viewed in pixels
- viewerHeight (num) - Height of image viewed in pixels
- viewerMag (num) - Magnification setting of the viewer when the data is saved
- task (factor) - Version number of platform
- inputFileName (chr) - File name of the input file

---

pilotHTT_RST *Simplified annotation data of the pilot study*

Description

This file is the aggregate of all clean data from the High-Throughput Truthing project. It has been cleaned of PII (names and emails) and other non-essential columns. This is the simplified version of 'pilotHTT' meant for the FDA Regulatory Science Tool (RST) Program.

Usage

```
pilotHTT_RST
```

Details

This data was collected from the CAmicroscope, PathPresenter, and eeDAP platforms. Please refer to for more information about the data.

As of 6 May 2022, this data contains 7898 observations of 10 variables.

Value

A data frame with the following (10) variables:

- batch (factor) - Batch number of image annotated by the reader (10 batches in total)
  - FDA-HTT-batch00x - Pilot Study annotations
  - FDA-HTT-Train00x - Expert Panel annotations
- WSI (factor) - Whole case file name of whole slide image annotated by reader
- caseID (factor) - ID for region of interest. Includes WSI, x position of ROI, y position of ROI, and length of ROI
- readerID (factor) - ID for each participant (profession with ID number)
  - There are four possible professions at the front end of readerID: pathologist, expert, resident, or unknown.
  - pathologist - board-certified pathologist
  - expert - member of the Expert Panel
  - resident - in residency
  - unknown - no indicated profession
- modalityID (factor) - Platform used by viewer (caMicro, pathPresenter, eeDAP, or camic-expert)
  - caMicro - Digital annotations collected
  - pathPresenter - Digital annotations collected
  - eeDAP - Microscope annotations collected
  - camic-expert - Digital Expert Panel annotations collected using the caMicroscope platform
- labelROI (factor) - Label of region of interest (Intra-Tumoral Stroma, Invasive Margin, Tumor with No Intervening Stroma, other regions)
- evaluable (logical) - Indicates whether the region of interest is appropriate for sTIL evaluation
- densityTILs (num) - Percent of area occupied by lymphocytes in Intra-Tumoral Stroma.
- experience (num) - Number of years of experience for pathologists. If experience == 100, experience is unknown
- experienceResident (num) - Number of years in residency for non-pathologists. If experienceResident == 100, experience is unknown

---

roisHTT *Information about the ROIs in this study*

Description

This file contains the information about the regions of interest (ROIs) in this pilot study. The data frame includes information about the image file names and the position of ROIs within the slides.

Usage

```
roisHTT
```

Details

There are 640 ROIs in the pilot study: 64 images x 10 ROIs per image. ROIs were selected before data collection. Please refer to this manuscript for information about ROI selection and to see a few samples.

Value

The data frame contains 10 columns:

- `task` (chr)- Task Pathologists were asked to complete
- `batch` (factor) - Batch Number
- `WSI` (factor) - Slide number
- `ROI` (factor) - Region of Interest analyzed
- `left`, `top`, `width`, `height` - All Numeric - Indicate the position of the ROI on the slide
- `widthMicrons` and `heightMicrons` - Numeric - Indicate the size of the ROI in Microns

---

`scannerInformationCasesHTT`

Scanner information of `casesHTT`

Description

This documentation includes the scanner information of `casesHTT`

Sample information

The pilot study slides and images were provided by key collaborators Roberto Salgado and Denis Larsimont (Chair Department of Pathology, Jules Bordet Institut). Support staff were Ligia Craciun (Lead at Tumorbank, Dr Science) and Sélim-Alex Spinette (Lab tech at Tumorbank).

The data include slides that are either ductal or lobular breast cancer cases. Differentiating between ductal and lobular is often obvious. In Belgian, these are denoted as CCI = carcinome canalaire (ductal) invasive and CLI = carcinome lobulaire (lobular) invasive.

The data include slides of matching (same patient) biopsies or surgical resections.

Biopsies are taken before surgery to make a diagnosis.

The slides shared were re-cuts. The original slides were imaged in 2017 and 2018. We have 115 images of the slides imaged in 2017, but we don't (yet) have the information to link these images to the re-cuts.

Slides come from FFPE blocks and H&E staining. Every digital image was checked. If the tissue was bent, the process was repeated as needed.

Slide 66: Two tumor nodes are described. The morphologies are similar, but we cannot define exactly which node was taken at biopsy.

Scanner information

Here are the details of the Nanozoomer 2.0-RS (Hamamatsu, Japan) under 40x magnification single-layer at the Jules Bordet Institute.

The images were scanned on one scanner, NanoZoomer 2.0-RS C10730 series. It is the high resolution and high-speed slide scanner that consists of Slide–feeder, X/Y Stage, Z focus motor, illumination system, optical components and TDI image sensor, this system realizes 1' 40" / (20x) per slide (20 mm x 20 mm scanning area). The NanoZoomer-RS system can automatically load up to

6 glass slides at a 20x or 40x magnification. All digital images were scanned in single layer at 40x magnification.

This scanner is equipped with a 3CCD-TDI camera which allow for brightfield and fluorescence images with only one camera. Resolution is 0.23um/px at 40x and 0.46um/px at 20x The system is equipped with a 20x, 0.75NA objective lens, with a 2x relay lens. The images are always acquired at optical 40x, we do a 2x2 binning on the camera to have 20x resolution This offers the advantage of keeping the depth of field of a 20x but with the resolution of a 40x. For fluorescence capabilities, the system can host up to 6 excitation filter, 2 dichroic mirrors, and 6 emission filters. Bordet is equipped to image Dapi/Fitc/Tritc/Cy3/Cy5 and equivalent.

You can download the NDP.view software for free from the Hamamatsu website.

Index