data-perparation-ammar-150454388

January 25, 2019

Table of Contents

0.1 Introduction

This report summaries the Data Preparation stage of the CRISP-DM cycle for this project. In particular, this report covers the background of the data and how it was modified for further analysis, modelling and evaluation of models. This report covers the methodology of data preparation not the detailed technical aspects - those are covered in the Sphinx documentation.

0.2 Data Background

As stated in the Business Understanding Stage report the data is provided from a kaggle competition in a dataset called the HAM10000. The dataset consists of a set of dermatoscopic images collected from various populations. The final dataset consists of 10015 images. Ground truths are provided by various confirmation techniques (follow-up examination, expert consensus or in-vivo confocal microscopy).

The data consists of the following csv files as alluded to in the business understanding report as well: * Images of skin lesions divided into two files: - HAM1000_images_part_1 - HAM1000_image_part_2 * HAM10000_metadata.csv: stores textual information about the image (ground truth, patient information, etc.) . * 28*28 Luminance and RGB values for the skin lesion images: - hmnist_28_28_L.csv - hmnist_28_28_RGB.csv

0.3 Final Dataset Description

0.3.1 Fields

- lesion_type (textual): The diagnosis (ground truth) as a textual description. Values:
  - Actinic keratoses

  - Basal cell carcinoma
  - Benign keratosis-like lesions
  - Dermatofibroma
  - Melanocytic nevi
  - Melanoma
  - Vascular lesions

- dx_type (textual): The method of diagnosis, textual. Values:
  - histopathology follow-up examination (follow_up)
  - expert consensus (consensus)
  - in-vivo confocal microscopy (confocal).

- lesion_type_idx: codes for diagnosis:
  - 0: Actinic keratoses
  - 1: Basal cell carcinoma
  - 2: Benign keratosis-like lesions
  - 3: Dermatofibroma
  - 4: Melanocytic nevi
  - 5: Melanoma
  - 6: Vascular lesions

- age (numeric): Natural numerical age of the individual the image is taken from.
- sex (textual): Sex of the individual the image is taken from (male, female or unknown).
- localization (textual): Location of skin lesion in individual.
- pixelXXXX_l_28_28 (numeric): Luminance value of images in 28 by 28 pixel representation.
- pixelXXXX_rgb_28_28 (numeric): RGB value of images in 28 by 28 pixel representation.

0.3.2 Table Preprocessing

Table processing was handled by src.data.make_datasetű script and tested by src.tests.test_make_dataset, both scripts are documented by the Sphinx documentation, hence this part of the report will avoid detailed technical explanations.

Metadata Cleaning Metadata cleaning involves the replacement of null/NA values with the average values of the age column. Averages are used instead of dropping to avoid losing the details these tuples carry, however this can effect the distribution of the data and needs to be taken into account later.

Moreover, new categorical numerical code and textual diagnosis fields are added for use in later stages, and the now not needed ids (lesion and image) are removed.

Luminance and RGB Values Cleaning The luminance and RGB datasets have their label fields removed since they aren't going to be used later in the analysis (already sorted)

Luminance and RGB Values Merge with Metadata The luminance and RGB pixel values (28 X 28) are added as predictors/variables to the metadata dataset for the Models. However, to ensure that their uniquely named a suffix (_rgb_28_28 or _l_28_28) for RGB or luminance pixels respectively. Lastly, the merge is carried via a column wise concatenation since no shared keys exists.

0.4 Summary

To summarise, the Data Preparation stage in this project forms the dataset for analysis, modelling (and their evaluation) using cleaning, merging and other methods to link and combine the raw data. These methods are carried in the src.data.make_dataset script and tested by src.tests.test_make_dataset script, both of which are covered in the Sphinx documentation.