business-understanding-ammar-150454388

January 25, 2019

Table of Contents

0.1 Report Introduction

This report documents the Business Understanding stage of the CRISP-DM process followed by this project (Chapman et al., 2000). This stage documents decisions from a business/stakeholder perspective, highlighting the objectives and requirements of the data mining project required to form its problem definition (the specification).

0.2 Organisation and Organisational Objectives

Even though this project is not directly related to any business or an organisation, this project needs to have a beneficial impact to problems faced by stakeholders that were not tackled before in the domain it covers.

0.2.1 Background

Pigment skin diagnosis field has been growing due critical importance of early detection of conditions like skin cancer. The field has developed various detection methods and the use of computerised algorithms to improve detection accuracy (Tschandl and Wiesner, 2018).

This project objective is to develop machine learning model for the diagnosis of pigmented skin lesions. The diagnosis method must be able to classify the each images to a given conditions (Actinic keratoses, intraepithelial carcinoma, etc.) (classification problem). The methods will also be evaluated using various model evaluation techniques (confusion matrices, F1 scores, etc.).

To achieve this, the project makes use of dermatoscopic images of pigmented skin lesions, which are normally collected for unaided eye diagnostic to train and tests machine learning models (Tschandl et al., 2018).

0.2.2 Resources and Data

The dataset provided is based on a dataset from a kaggle competition ("Skin Cancer MNIST," n.d.) (Tschandl et al., 2018). The dataset is called the HAM10000 ("Human Against Machine with 10000 training images") dataset, which is a set of dermatoscopic images collected from a diverse populations. The final dataset consists of 10015 images whose ground truths (skin lesions) are confirmed through histopathology follow-up examination (follow_up), expert consensus (consensus), or by in-vivo confocal microscopy (confocal). This dataset was formed to attempt to resolve the issue of small sample sizes and lack of diversity of the skin lesion datasets ("Skin Cancer MNIST," n.d.).

The provided datasets consist of ("Skin Cancer MNIST," n.d.): * Images of skin lesions divided into two files: - HAM1000_images_part_1 - HAM1000_image_part_2 * HAM10000_metadata.csv: stores textual meta information about the image (ground truth, patient information, etc.) . * *8 and 28*28 Luminance and RGB values for the skin lesion images: - hmnist_8_8_L.csv - hmnist_8_8_RGB.csv - hmnist_28_28_L.csv - hmnist_28_28_RGB.csv

0.2.3 Success Critera and Objectives

To obtain what criterion should the models pass a look into what stakeholders on the field expect is necessary. To do this, a look into published work and resources in the area is necessary.

As stated in the kaggle page, the dataset was only setup for a selection of conditions ("Human Against Machine with 10000 training images"): * Actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec) * basal cell carcinoma (bcc) * benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, bkl) * dermatofibroma (df) * melanoma (mel) * melanocytic nevi (nv) * vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc).

Any solution model needs to be setup to classify the skin lesions to these conditions that the dataset was setup for.

Also as previously stated in the background, early diagnosis is critical, so minimising false negatives should be a clear priority (improving f1 scores). When it comes to the accuracy of Machine Learning methods for pigmented skin diagnosis, in (Dreiseitl et al., 2001), discriminatory or classification performance was only considered. (Dreiseitl et al., 2001) has also found that decisions trees do not translate well in this field and the results from logistic regression, neural networks and support vector machines were the best. On the other hand (Gautam et al., 2018) found that SVM also performed really well but also Random Forest seemed to top the accuracy of the tested models.

However, both papers seem to miss non-discriminatory evaluation methods, concentrating on accuracy but not other factors (e.g. computation), which might become important since the importance of early diagnosis might mean large datasets from the population need to be processed quickly as early scanning becomes encouraged. Moreover, as stated in the background, one of the core issues of machine learning problems in this area is the lack of diversity in the data which shows in (Dreiseitl et al., 2001) and (Gautam et al., 2018) as the papers only classified 3 conditions and melanoma detection only respectively.

Therefore based on the background reading, the following criteria needs to be met:

- Develop a neural, logistic, SVM and/or random forest models
- Train the models using a diverse and large sample sized dataset
- Classify the diagnosis based on a selection of skin conditions (stated above)
- Find and analyse model confusion matrices and measure models accuracy (discriminatory)
- Find and analyse model fit training speed and test/usage speed (non-discriminatory)

- Select a favourable model based on the evaluation measures, but prioritise good false negative performance

0.3 Project Specification and Plan

0.3.1 Requirements and Rationale

- Convert images/datasets to a form that can be classified and clean the data
  - Reasoning: Prepare the data for classification by recording predictors and remove data that might distract the analysis
- Merge all classification related clean dataset together
  - Reasoning: create the final dataset for later stages by collecting and aggregating required fields
- Perform Exploratory Data Analysis
  - Reasoning: Gives clue about data structure and relationship to help with fitting and analysing the models
- Fit SVM, RF, LR and Neural Network models and measure fitting times
  - Reasoning: Fit using training data to test for evaluation measurements
- Test and analyse models for accuracy, error types (confusion matrix) and computation times
  - Reasoning: Used to collect the required information for the final analysis and model selection

0.3.2 Key Tools

Analytics and modelling steps will be performed using python code and libraries (e.g. matplotlib, sklearn, etc), which are tested with the pytest library and reported on using python notebook (literate programming). Documentation for the code will be managed using reStructuredText based Sphinx and saved as HTML pages ("Overview — Sphinx 2.0.0+/4f37b33 documentation," n.d.). The project structure template is created using a Cookie Cutter Data Science hierarchy ("Home - Cookiecutter Data Science," n.d.) which provides a Cookie Cutter and Python based environment to work with. the entire project hierarchy is version controlled using git and GitHub.

0.3.3 Project Plan and Risks

As stated in the introduction to this report, the project will follow the CRISP-DM hierarchical cyclic process model for data mining. The process is typically done iteratively in order, but flexibility is possible and will be used to carry Data Preparation before Understanding to ensure that the datasets/images are processed for the EDA (Chapman et al., 2000).

Test Driven Development will be carried using PyTests and literate programing (Jupyter Notebook) will be used for the analysis documentation. These tools ensure that both tasks can be carried simultaneously with the code, which reduces the risk of testing and documentation issues when they are rushed when time constraints occur.

GitHub is used for version control to ensure that any risk of changes causing behaviours that are difficult to undo are minimised through version control. Moreover, its use ensures that the

project development route can branch and change easily without worry over the risk of not being able to revert back when the changes are scrapped.

0.4 References

1. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. CRISP-DM 1.0 Step-by-step data mining guide.
2. Gautam, D., Ahmed, M., Meena, Y.K., Ul Haq, A., 2018. Machine learning–based diagnosis of melanoma using macro images. International Journal for Numerical Methods in Biomedical Engineering 34. https://doi.org/10.1002/cnm.2953 3.Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data 5, 180161. https://doi.org/10.1038/sdata.2018.161
3. Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., Binder, M., 2001. A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions. Journal of Biomedical Informatics 34, 28–36. https://doi.org/10.1006/jbin.2001.1004
4. Home - Cookiecutter Data Science [WWW Document], n.d. URL https://drivendata.github.io/cookiecutter-data-science/ (accessed 1.4.19).
5. Overview — Sphinx 2.0.0+/4f37b33 documentation [WWW Document], n.d. URL http://www.sphinx-doc.org/en/master/ (accessed 1.4.19).
6. Skin Cancer MNIST: HAM10000 [WWW Document], n.d. URL https://kaggle.com/kmader/skin-cancer-mnist-ham10000 (accessed 1.4.19).
7. Tschandl, P., Wiesner, T., 2018. Advances in the diagnosis of pigmented skin lesions. British Journal of Dermatology 178, 9–11. https://doi.org/10.1111/bjd.16109
8. Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data 5, 180161. https://doi.org/10.1038/sdata.2018.161