written-report-ammar-150454388

January 25, 2019

Table of Contents
1 Introduction
2 Project Background and Justification
3 Project Criteria/Objectives
4 Tools, Methodology and Evaluation Design
5 Exploratory Findings and Results
6 Evaluation and Future work

0.1 Introduction

This report covers the Machine Learning data mining project on the diagnosis of skin lesions by summarising the project's background, objectives, methods and findings. Moreover, this report will also reflect on the project processes and tools. This report however is a summary of the CRISP DM documents produced by this project, hence detailed findings will not be covered here.

0.2 Project Background and Justification

Pigment skin diagnosis field has been growing due to the importance of early detection of skin conditions like skin cancer. The field has developed various detection methods and the use of computerised algorithms for accuracy. Hence, the use of modelling techniques to improve that rely on dermatoscopic images of pigmented skin lesions for training cannot be understated.

0.3 Project Criteria/Objectives

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 set of given conditions. The models will also be evaluated using various model evaluation techniques (confusion matrices, F1 scores, etc.) as part of the project. The specific criteria that this project needs to complete went as below:

- 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

0.4 Tools, Methodology and Evaluation Design

The project will follow the CRISP-DM hierarchical cyclic process model for data mining (Chapman et al., 2000) to provide the process for data mining in this project. The project and make use of Test Driven Development (PyTests) and literate programing (Jupyter Notebook), which are used to ensure that both tasks of testing documentation can be carried simultaneously with the coding process. Git and Github is used for version control to manage the progress of cookie cutter based project template, the code in that project will also be documented with the help of Sphinx.

The project will be divided into these stages according to the produced CRISP DM based reports:

1. Business Understanding:
    - This report covers the background, justification, objectives and project plan for the data mining project.
2. Data Preparation:
    - This report covers how the data was cleaned and prepared.
3. Data Understanding:
    - This report covers the data exploratory process undertaken in this project
4. Modelling and Evaluation:
    - This report covers the construction of the machine learning models (logistic, SVM and random forest models) and their evaluation. This report also covers the evaluation of the overall project.

0.5 Exploratory Findings and Results

The exploratory analysis of the data has shown that skin lesions are more likely to occur at extremities, the back and trunk. Also, lesions are overwhelmingly of the Melanocytic Nevi type, and that lesions have varying degrees of dependence on age with some getting effected less than others.

When it comes to models, their classification performance (f1 and accuracy) has shown that logistic regression seems to best the two other models, with the the SVM model particularly performing worse. For run times, the Random forest was slow to fit and the SVM model was very slow to predict compared to the other two. Overall, the logistic model seemed to be significantly faster than the two overall according to the combined fit and prediction times.

0.6 Evaluation and Future work

While the results show that out of the three models the logistic regression is best, the overall f1 scores and confusion matrices were poor when it came to their consistency in all the models. This is perhaps caused by logistic regression's poor dimensionality performance and poor tuning for the SVM and Random Forest, as the SVM was not tuned due to the use of a Bagging and One Vs All classifier, and the Random Forest had a narrow set of tuning parameters due to run time issues. Moreover, as stated in the exploratory findings, the lesions are overwhelmingly of the Melanocytic Nevi type which might have made it difficult to balance the tuning/fit and to interpret the f1 scores.

Hence, in future projects, when solving problems with high dimensionality it might be wise to use a method that is more paralyzable (e.g. Neural Networks) to make tuning easier with parallel processing based Graphics Cards. Moreover, it might be wise to make use of dataset that is not overwhelmingly dominated by one class.

When it came to the techniques and tools used in this project, while the newer cookie cutter python based environment provided useful tools with more intricate controls (e.g. VirtualEnv for isolation), it proved to be a significant hurdle in the early stages of the project. The newer tools seemed to not have as many conveniences like some R based tools like Project Template (e.g. automatic dataset loading). Nonetheless, the experiences with some of the tool - in particular Sphinx documents and virtual enviroments - will prove useful in future projects.