Logistic Regression on the BIOSCAD data set

Introduction

**Overview:**

The aim of the project was to create a model that, based on biomarkers measured on the skin, would diagnose atopic dermatitis (AD) in a patient. We used the data from the AMC BIOSCAD study.

Method

**Data set:**

The AMC BIOSCAD study data is first pre-processed to make it suitable for training – see pre-processing report (../Preprocessing/preprocessing-report.pdf) for details. One additional step was performed to the pre-processed data – the SCORAD values were replaced with a Boolean to indicate if the patient suffered from AD. After pre-processing, there were 100 observations (53 AD and 47 control) available for training, 35 features and one dependent variable.

**Model building:**

Building each model consisted of a training, validation, and testing phase. To train the model, 60% of the data was used. Using the validation data (20%), the best prediction threshold was found. This is the value above which a patient is predicted to have AD, and below which not. Once the best prediction threshold has been found, the data is tested using the remaining 20% to allow the performance of the model to be evaluated. This 3 step process is repeated 100 times using different data points for the training, testing, and validation data.

**Performance evaluation:**

Several performance metrics were calculated to test the models. Since the two classes are balanced in the dataset, our main metric is the accuracy – the percentage of predictions that were correct. We also calculated the sensitivity (true positive rate, TPR), and the specificity (true negative rate, TNR). We also computed the F1-score, the Matthews correlation coefficient and the diagnostic odds ratio.

**Input variable selection:**

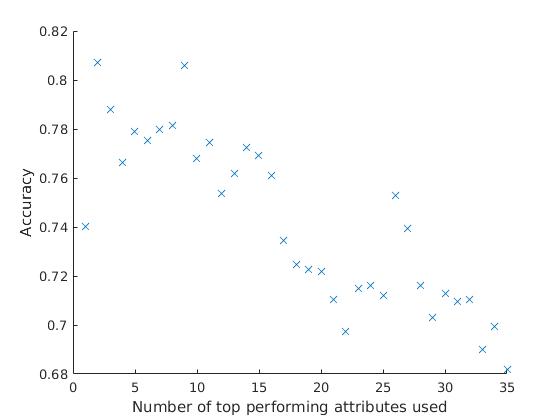
The first model built used all 35 attributes as inputs. To see the relative importance of these attributes, we also built models using each attribute individually. Using this information, we combined the top *n* performing variables (based on accuracy) together to create input data for new models.

Figure 1

As increasing the number of top performing attributes used decreases the performance of the model (figure 1), we chose to limit future models to 4 input attributes. We then built every possible model using up to 4 attributes (approximately 35,000 models).

We also performed logistic regression using only IL-1a and IL-1β as these were the only two attributes in which most of data was above the detection limit.

Results

**Top performing models:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute 1** | **Attribute 2** | **Attribute 3** | **Attribute 4** | **Accuracy** | **Threshold value** |
| FLGCarrier | IL-7 | IL-2 | - | 0.8365 | 0.4109 |
| FLGCarrier | IP-10 | IL-7 | IL-2 | 0.8350 | 0.4314 |
| FLGCarrier | MCP-1 | IL-7 | IL-2 | 0.8345 | 0.4051 |
| FLGCarrier | IL-7 | IL-2 | IL-13 | 0.8340 | 0.4340 |
| FLGCarrier | MCP-4 | IL-7 | IL-2 | 0.8335 | 0.3768 |
| FLGCarrier | IL-7 | IL-16 | IL-2 | 0.8330 | 0.4583 |
| FLGCarrier | IL-7 | IL-2 | IL-17A | 0.8315 | 0.4332 |
| FLGCarrier | MIP-1A | IL-7 | IL-2 | 0.8310 | 0.3664 |
| AgeAtVisit | FLGCarrier | IL-7 | IL-2 | 0.8305 | 0.4072 |
| FLGCarrier | IL-7 | IL-2 | IL-6 | 0.8305 | 0.4172 |
| FLGCarrier | IL-7 | IL-15 | IL-2 | 0.8270 | 0.3980 |
| FLGCarrier | IL-16 | IL-2 | IL-10 | 0.8270 | 0.3811 |
| skinType6 | IL-5 | IL-16 | IL-2 | 0.8270 | 0.4102 |
| IL-5 | IL-7 | IL-16 | IL-2 | 0.8255 | 0.4214 |
| FLGCarrier | IL-7 | IL-18 | IL-2 | 0.8250 | 0.4014 |

Table 1

**Most accurate model evaluation:**

|  |  |
| --- | --- |
| **Performance metric** | **Value** |
| Accuracy | 0.837 |
| Sensitivity | 0.891 |
| Specificity | 0.779 |
| Precision | 0.817 |
| F-Score | 0.852 |
| Mathews Correlation Coefficient | 0.677 |
| Diagnostic odds ratio | 28.940 |

Table 2

**Most accurate model coefficients:**

|  |  |
| --- | --- |
| **Attribute** | **Odd ratio** |
| IL-2 | 0.287 |
| IL-7 | 0.328 |
| FLGCarrier | 39.425 |

Table 3

The coefficients of the best model (table 3) show that FLG-Carrier had the most important contribution for the prediction. This is because in the original data set, there are no patients with FLG mutations in the control group. As such, the model indicates that any patient with an FLG mutation must have AD. However, we know from other analysis of other data sets that this is not the case. As a result, this variable is overfitting the input data and will not generalise well to external data. To avoid this problem, we looked at models that did not use FLGCarrier as a feature.

**IL-1a and IL-1β model evaluation:**

|  |  |
| --- | --- |
| **Performance metric** | **Value** |
| Accuracy | 0.541 |
| Sensitivity | 0.575 |
| Specificity | 0.505 |
| Precision | 0.561 |
| F-Score | 0.568 |
| Mathews Correlation Coefficient | 0.080 |
| Diagnostic odds ratio | 1.378 |

Table 4

Discussion

**Input attributes:**

The most accurate model used 3 attributes as inputs. These were FLG-Carrier, IL-7, and IL-2. The 10 next most accurate models also used these 3 inputs and one additional input. This indicates that these 3 attributes have the largest effect on whether or not a patient is predicted to have AD.

**Reliable data model:**

The model trained on only the reliable data (IL-1a and IL-1β) showed poor accuracy levels (table 4). These were not much better than a model which would randomly guess either AD or not.

**Input data quality:**

Another issue is that the majority of data in the BIOSCAD data set is marked as below the detection range. As such, the data is unreliable and thus the models also unreliable (garbage in, garbage out principle). This is shown as the best selected continuous attributes (IL-2, and IL-7) have already been shown to have limited impact on SCORAD.