CAPSTONE PROJECT – STATISTICAL DATA MINING FOR BIG DATA

**Using Machine Learning Methods to Detect Autism Spectrum Disorder in Adults**

**William Belcher**

Autism Spectrum Disorder is a neurodevelopment condition in which the people with the condition exhibit a variety of distinct behavioural patterns. Diagnosis typically happens at a young age as parents and teachers notice the autism symptoms. However, identifying autism in adults is particularly difficult as its symptoms overlap with a variety of other mental health conditions. With machine learning techniques growing ever popular in the clinical setting, a rapid screening process that could be used in assisting the referral of patients to medical professionals was essential. In this paper, the possibility of using Logistic Regression, Naïve Bayes and tree-based classifiers is explored. The proposed methods were evaluated on a publicly available AQ-10-Adult based screening data set that contained 704 observations of 21 attributes. After the relevant pre-processing of the data and application of the above-mentioned techniques, the achieved results indicated that a classification tree on a reduced data set containing the first two principle components derived from the AQ-10 questionnaire and a subset of the environmental predictors

**Introduction**

Autism Spectrum Disorder (ASD) is a neurological condition associated with many atypical mannerisms and behavioural patterns, most notably those surrounding interpersonal interactions. Autism Spectrum Disorder is a condition related to the development of the human brain. It is worth noting that both environmental and genetics may be contributing factors in the development of ASD. However, scientists have been unable to uncover the root cause and as such ASD is usually detected through observations and diagnosed by a specialist. Unfortunately, the process for receiving an ASD diagnosis are lengthy with multiple appointments with specialists not being cost effective. Early detection of the condition can assist in the improvement of the subject’s overall health by enabling them to implement techniques and medication that reduce the impact of their condition on their daily lives sooner. With the rapid increase in modern computing power and number of machine learning models assisting in the diagnosis of medical conditions, the early detection of ASD based on variety of physiological attributes now seems viable. The detection of autism spectrum disorder in a patient proves difficult as, as the name implies, the disorder is a spectrum resulting in significant intragroup variance in those being classified as having the condition. A time-efficient and easily accessible screening process is necessary in assisting medical professionals in informing individuals whether they should pursue a formal, clinical diagnosis.

**Data**

The Autism Screening Data for Adults data set, collected from the UCI Machine Learning Repository, contains several predictors and one target variable [https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult]. The attribute types for these predictors are either categorical, continuous or binary with the response variable being a categorical “yes” or “no”. The dataset’s 20 attributes used for prediction are listed below:

|  |  |
| --- | --- |
| **Attribute ID** | **Description** |
| 1-10 | Answer to the corresponding AQ-10-Adult [] question |
| 11 | Patient age |
| 12 | Gender |
| 13 | Ethnicity |
| 14 | Did the patient have jaundice at birth? |
| 15 | Family history of ASD |
| 16 | Patient country of residence |
| 17 | Has the patient used the screening app before? |
| 18 | Screening score |
| 19 | Age group |
| 20 | Person who’s using the screening app’s relation to patient |

Table 1: Attribute list

In total, there were 704 observations. It is worth noting that the screening score is the sum of the answers to the AQ-10 questions, where a 1 represents a ‘slightly agree’ or ‘strongly agree’ and a 0 represents a ‘slightly disagree’ or ‘strongly disagree’.

*Data Pre-processing*

Data pre-processing is a technique in which the raw data is transformed into a meaningful and understandable format. The ‘Age group’ column was removed from the data as it only contained one value of ’18 and over’. The ‘Screening score’ column was also removed as it was just the sum of the answers to the AQ-10 questions. The “Used screening app before” as well as the “relation to patient” columns were removed as they are deemed unnecessary by inspection. The “country of residence” was also not included in the model training data as it was deemed unimportant for the scope of this task. Missing values in the data are denoted with a ‘?’ with rows containing missing values being removed from the data. Missing value imputation was not conducted in order to not introduce unnecessaryvariance in the predictors. A single outlier was detected and removed. The “patient ethnicity” column contained 11 categories. In order to reduce the dimensionality of the data, these categories were collapsed in 4, more general, categories being: “White”, “Asian”, “Black”, “Other”. The size of the cleaned data used for model generation contained 17 columns, 16 predictors and 1 response, with 608 total observations.

**Methods**

*Principle Component Analysis (PCA)*

Principal component analysis is an unsupervised learning technique used for dimension reduction as well as exploratory data analysis by projecting the variables onto a new, orthogonal basis that can be used to illustrate the proportion of variance explained by each principal component.

*Training and testing split*

The complete data set was split into training and testing subsets using an 80/20 split. K-fold validation with the training subset and with k = 5 was used for each model. Figure 1 shows the final training, validation and testing sets that were used.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Complete Dataset | | | | | | |
|  | | | | | | |
|  | | | | | | |
| Training Set (80%) | | |  | Testing Set (20%) | | |
|  | | | | | | |
|  | | | | | | |
| 1 | 2 | 3 | | | 4 | 5 |

Figure 1: Data split

*Logistic Regression (LR)*

Logistic regression is based off the standard linear regression methods but applies the logit transformation to the resulting in a response that will lie between the value of 0 and 1 inclusive.

*Naïve Bayes (NB)*

The naïve Bayes classifier is a probabilistic machine learning model that is based on Bayes theorem, finding the probability of happening given that has already occurred. The naivety of the model comes from the assumption that features are independent in each class.

*Classification Tree (CT)*

Classification trees are a subtype of decision trees where the target variable can only take a discrete set of values. In the tree structures, the leaves represent the class labels with the branches representing conjunctions of features that lead to those class labels.

*Random Forest (RF)*

The random forest algorithm is an extension of the classification tree that fits many classification trees to a data set and the combines the predictions from all of the trees. [https://www.sciencedirect.com/topics/computer-science/classification-tree / Disease Modelling and Public Health, Part A]

*Best Subset Selection*

Best subset selection is a method that aims to find the subset of independent predictors that best predict the response and does so by comparing all possible combinations of the predictors. This method works well for data with small dimensions, but as the predictors increase linearly, the possible number of combinations increases exponentially.

**Results and Discussion**

*Performance Metrics*

The result of a model is measured in terms of its specificity, accuracy and sensitivity. The values are obtained by using the generated models to predict the outcome on the test data set and making note of the resulting confusion matrix.

|  |  |  |
| --- | --- | --- |
|  | Predicted Outcome | |
| Actual Outcome | True Negative | False Positive |
| False Negative | True Positive |

Table 2: Confusion Matrix

|  |
| --- |
| Performance Metrics |
|  |
|  |
|  |

Table 3: Performance Metrics formulae

Principle component analysis was conducted on the AQ-10-Adult questions, in order to find the minimal number of questions required to explain a significant amount of the variance in the data. PCA illustrated that there was no significant ‘elbow’ in the proportion of variance explained by each principal component. However, with only 3 components, approximately 50% of the variance in the response could be explained as seen in Figure [].

Chart

Description automatically generated

These three components, along with the patient’s age, gender, collapsed ethnicity, family history of autism and whether they suffered from jaundice as a child were used in training the following models.

Best subset selection was first completed on the logistic regression model. This returned a model which only contained the three principle components and the ethnicity. Of these four predictors, only PC1 and PC3 were found to be statistically significant. Creating an LR model with the response being predicted purely by PC1 and PC3 achieved an accuracy of 99.18% on the test set. This high accuracy from only two principle components can be attributed to the way ASD is diagnosed. Figure **X** illustrates the data projected onto the first two principle components and reveals distinct clusters exposed from PCA.

Chart, scatter chart

Description automatically generated

A psychiatrist will make their diagnosis after looking exclusively at the behavioural patterns displayed by the patient which the AQ-10-Adult questionnaire aims to identify. The psychiatrist does not consider other aspects of the patient such as their age, gender, etc.

For completeness’ sake, the full training data was used for constructing the remaining models. Figure **X** illustrates the range of cross validated error produced from each model:

Chart, box and whisker chart

Description automatically generated

From the figure, it can be seen that the classification tree proved most stable whilst also having the highest mean cross validated accuracy of 98.76%.

All models were then tested with the 20% test set. Table **X** provides the full results from all the models tested:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Sensitivity | Specificity |
| LR | 99.18% | 100% | 97.50% |
| NB | 95.59% | 97.72% | 95.40% |
| CT | 100% | 100% | 100% |
| RF | 97.73% | 96.42% | 98.26% |

The testing set corroborates the cross validation error estimate.

**Conclusion**

**References**

* [4] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA <http://www.rstudio.com/>

Mode: Desktop. Version 1.25042