CLASSIFICATION OF AUTISTIC SPECTRUM DISORDER SCREENING DATA

IST 707 : FRANCISCO NUNEZ-FONDEUR

06/12/19

Problem

- Autism Spectrum Disorder (ASD) affects an estimated 1 out of 59 children in the US.
- The process to get properly diagnosed was lengthy and expensive and new screening tools were needed.
- The Autism Spectrum Quotient (AQ) screening questionnaire was created consisting of 50 questions to be administered by a medical professional. A condensed version (AQ-10) for self diagnosis was created consisting of 10 questions. The AQ-10 does not give a definitive diagnosis, but a score > 6/10 indicates that you may have ASD and should seek a professional diagnosis.
- Although the AQ-10 tool exists, there aren't many available datasets associated with clinical screenings and behavior, most are genetic in nature. More data on screenings and behavior would lead to improving existing tools and make it faster/easier to diagnosis ASD
- The Goal is to classify some of these behaviors/characteristics

Data

- Three separate data sets for children age 4-11 (292), 12-17 (104), and adults 18+ (704) from the UCI Machine Learning repository
  - https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++#
  - https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Adolescent+++
  - https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult
- The three sets were combined into one data set

Data

Table 1: Features and their descriptions

Attribute	Type	Description
Age	Number	Age in years
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician, etc.
Country of residence	String	List of countries in text format
Used the screening app before	Boolean (yes or no)	Whether the user has used a screening app
Screening Method Type	Integer (0,1,2,3)	The type of screening methods chosen based on age category (0=toddler, 1=child, 2= adolescent, 3= adult)
Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner

Data

- Testing/Train Data would be created and classification would be performed using various algorithms.
  - Decision Tree (c5.0)
  - Naïve bayes
  - K-NN
  - Support Vector Machine (SVM)
- Data had to be prepped. There were a few NA values.
  - Age NA were replaced with the mean age for each data set (child, adolescent, adult)
  - For discrete values an "unknown" value was created (ex: unknown ethnicity, unknown country of origin)
  - No observations dropped. 1100 total observations

Initial Results

- The AQ-score was a strong classifier. Every classification method was able to predict an ASD outcome 100% of the time based on this value. Clearly something was up.

```
Evaluation on training data (727 cases):

           Decision Tree
         ----------------
         Size      Errors

            2    0( 0.0%)   <<


          (a)   (b)    <-classified as
         ----  ----
         467           (a): class NO
               260     (b): class YES


        Attribute usage:

        100.00% result


Time: 0.0 secs
```

```
Total Observations in Table:  373


                     | predicted diagnosis
actual diagnosis     |        NO |       YES | Row Total |
---------------------|-----------|-----------|-----------|
                  NO |       240 |         0 |       240 |
                     |     0.643 |     0.000 |           |
---------------------|-----------|-----------|-----------|
                 YES |         0 |       133 |       133 |
                     |     0.000 |     0.357 |           |
---------------------|-----------|-----------|-----------|
        Column Total |       240 |       133 |       373 |
---------------------|-----------|-----------|-----------|
```

Pruning Data

- The AQ-10 scores and answers were removed to focus on the other attributes
- According to Autism Speaks, an advocacy group in the US, ASD is not prevalent in any specific ethnic group, socio-economic class, or country of origin. As a result the following variables were dropped
  - Ethnicity – 12 factors
  - Country of Origin – 89 Factors
  - Age and Age Category were measuring the same thing. Age was dropped and Age Category was kept as a factor

Decision Tree

- C5.0 Algorithm (c50 package) created a decision tree based on Age Category and Relationship of the person who filled out the survey.
- The tree was able to predict 68.63% of outcomes. Classifier was not very accurate.
- While it has an 80.83% of correctly identifying a "No-ASD" classification. There is low accuracy with classifying a true "Yes-ASD" (46.61%)

```
C5.0 [Release 2.07 GPL Edition]        Wed Jun 12 03:09:34 2019
-------------------------------

Class specified by attribute `outcome'
Read 727 cases (18 attributes) from undefined.data

-----  Trial 0:  -----

Decision tree:
12to17 > 0: YES (68/26)
12to17 <= 0:
:...18Plus > 0: NO (473/129)
    18Plus <= 0:
    :...RelationRelative > 0: NO (12/3)
        RelationRelative <= 0:
        :...RelationUnkown <= 0: YES (141/66)
            RelationUnkown > 0: NO (33/11)
```

```
*** boosting reduced to 2 trials since last classifier is very inaccurate
*** boosting abandoned (too few classifiers)


Total Observations in Table:  373

                 | predicted diagnosis
actual diagnosis |        NO |       YES | Row Total |
-----------------|-----------|-----------|-----------|
              NO |       194 |        46 |       240 |
                 |     0.520 |     0.123 |           |
-----------------|-----------|-----------|-----------|
             YES |        71 |        62 |       133 |
                 |     0.190 |     0.166 |           |
-----------------|-----------|-----------|-----------|
    Column Total |       265 |       108 |       373 |
-----------------|-----------|-----------|-----------|
```

Naïve Bayes

- Naïve Bayes applied using the e1071 package. Using a laplace value of 3
- The classifier performed better since it was not looking at interdependence. It had a 96.51% accuracy.

actual diagnosis	predicted diagnosis NO	YES	Row Total
NO	230 0.958 0.987	10 0.042 0.071	240 0.643
YES	3 0.023 0.013	130 0.977 0.929	133 0.357
Column Total	233 0.625	140 0.375	373

K-NN

- K-NN algorithm was applied using the class package
- Without scaling, there was an accuracy of 62.4% . After scaling, the accuracy only improved to 68.63%. Its worth noting that after, there was a significant improvement in positively identifying No-ASD however there was a large number of false negatives and Yes-ASD was not categorized properly.

	knnAutSpread		
testAutlabel	NO	YES	Row Total
NO	186	54	240
	0.775	0.225	0.643
	0.684	0.535	
	0.499	0.145	
YES	86	47	133
	0.647	0.353	0.357
	0.316	0.465	
	0.231	0.126	
Column Total	272	101	373
	0.729	0.271	

	knnAutPrediction		
Total2z_test_label	NO	YES	Row Total
NO	247	26	273
	0.905	0.095	0.744
	0.744	0.743	
	0.673	0.071	
YES	85	9	94
	0.904	0.096	0.256
	0.256	0.257	
	0.232	0.025	
Column Total	332	35	367
	0.905	0.095	

SVM

- SVM algorithm was applied using the kernlab package
- The use of the vanilla kernel had a 65.68% accuracy. The other kernals did not improve the accuracy. Rbfdot (Radial basis/Gaussian) performed the best out of the other kernel with a 64.87% accuracy. Again the big issue is false negatives.

Linear (vanilla) kernel function.

Number of Support Vectors : 504

Objective Function Value : -491
Training error : 0.337001

actual number	predicted number NO	YES	Row Total
NO	220 0.590	20 0.054	240
YES	108 0.290	25 0.067	133
Column Total	328	45	373

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma = 0.04796701162361 42

Number of Support Vectors : 510

Objective Function Value : -465.9183
Training error : 0.295736

actual number	predicted number NO	YES	Row Total
NO	183 0.491	57 0.153	240
YES	74 0.198	59 0.158	133
Column Total	257	116	373

Results

Including AQ-10 Reponses & Score		
Model	Predicted Accuracy %	Actual Accuracy %
Decision Tree	100.00%	100.00%
Naïve Bayes	100.00%	100.00%
k-NN	100.00%	100.00%
SVM	100.00%	100.00%
Excluding AQ-10 Reponses & Score		
Model	Predicted Accuracy %	Actual Accuracy %
Decision Tree	67.00%	68.64%
Naïve Bayes	98.93%	95.98%
k-NN	68.09%	69.75%
SVM	68.09%	68.09%

Conclusion – What possibly went wrong?

- Data
  - Not enough data
  - Bias results. In the actual data, 100% of the AQ-10 results correctly predicted the diagnosis. According to a study in *Phsychological Medicine* (Ashwood, etal 2016), AQ-10 scores are not significant predictors of ASD and 64% of those who scored below the cutoff were later identified as false negatives. The actual data had no false negatives in line to what would be expected in this study
- Complexity of ASD
  - ASD is still being studied. While there have been some improvements in diagnosis and understanding of this condition, it is still not fully understood and I think there are other variables at play which weren't record as part of this set.
  - More Data should be collected, and they should look at more potential variables.

* Ashwood, K. L., Gillan, N., Horder, J., Hayward, H., Woodhouse, E., McEwen, F. S., . . . Murphy, D. G. (2016). Predicting the diagnosis of autism in adults using the autism-spectrum quotient (AQ) questionnaire. Psychological Medicine, 46(12), 2595-2604. doi:10.1017/S0033291716001082