Course Project

Due Dec 6, 2019

1 Introduction

This project requires you to explore classification algorithms on real-world datasets, and write a report explaining your experimental results. The language of implementation is up to you — the requirements are: 1) your program be able to accept the data files (posted with the project assignment in BB) as the input to your program, and 2) your program be able to classify instances and analyze risk factors. You are free to construct whatever user interface for your program, but you must *fully document* your interface.

2 Algorithm

- **Your algorithm should be based on algorithms and techniques learned in our course**. Usually a straight forward implementation of one method will not lead to satisfactory performance. Your algorithm can be a combination of methods and should incorporate one or more data mining techniques when the situation arises. These techniques may include (and certainly not limited to):
  - Handling imbalanced dataset
  - Proper imputation methods for missing values
  - Different treatment of various type of features: continuous, discrete, categorical, etc.

3 Data

You'll be examining the behavior of your classification algorithm(s) on Autism Spectrum Disorder (ASD) screening for adults, adolescents and children. The three datasets are obtained from the UCI machine learning lab. Each dataset is presented with a description file and a data file.

Each data file contains 20 features to be utilized for predicting the ASD cases and determining the influential autistic traits. In particular, there are ten behavioral features based on answers to 10 questions, plus ten individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behaviour science. The 10 questions are designed differently for the adults, adolescents and children groups. Therefore, you need to be careful if you plan to combine these datasets for further analysis. Questionnaires for each group are also posted in BB for your reference.

4 Goal

The Goal of the project is to accurately predict the ASD cases and analyze what are the most important factors in ASD prediction.

5 Your Mission...

Deliverables for this project are:

- Code to implement the classification algorithm(s) for the given datasets
- **A README file, with simple, clear instructions on how to compile and run your code**
- A discussion of data mining techniques employed in your algorithm
- Testing statistics of your learning algorithm(s). At a minimum you should provide training and test set accuracies
- Analyze risk factors of ASD identified by your model
- A report analyzing the behavior of your algorithm on the datasets, including any unusual or anomalous (in your opinion) behavior

6 How to turn in your code

- **Your program must run on erdos.dsm.fordham.edu**
- **Zip all your files (code, README, written report, etc.) in a zip file named** $\{firstname\}\_\{lastname\}\_CS6930\_project.zip$ **and upload it to Blackboard**
- **Only one person in your group needs to turn in the code and the report. Make sure every team member's name is listed on the cover of the report**