Status Report: Predicting progression of Alzheimer's Disease with clinical and genotype data

Abstract

Machine learning algorithms have the potential to predict Alzheimer's disease (AD) progression by analyzing large clinical and genomic datasets. Here, we describe our progress on our implementation of ensemble methods to generate accurate predictions from a large AD database. We are working with the data from 767 patients (split into training and testing sets), with plans to supplement our data with work from additional longitudinal studies. We have plans to gain domain expertise from the Penn doctors who originally developed the database we have accessed.

1. Introduction

Alzheimer's disease (AD) is predicted to affect 1 in 85 people globally by 2050, causing dementia and eventual death. Care in the US costs $100 billion annually, and the available drugs can only help relieve some symptoms (Duthey, 2013).

1.1. Motivation

It is currently difficult to predict the progression of AD, and it often progresses undiagnosed for years. Machine learning algorithms have the potential to assist doctors and patients by accurately predicting disease progression based on clinical and genetic data, which would enable accurate, early diagnoses.

1.2. Related work

Since the causes of AD are currently unknown and there are no laboratory tests that can accurately perform a diagnosis, AD progression is quantified with psychological tests like the mini-mental state examination (MMSE) - a questionnaire used to measure cognitive impairment. This set of 30 questions was developed in 1975 and remains the standard (Doerflinger, 2007)

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Machine learning algorithms have been used on ADNI data with varying success to predict the change in MMSE. Interestingly, no single algorithm has been shown to be superior across all AD datasets, particularly when progression is measured up to varying time points (Umer, 2011)

2. Materials & Methods

2.1. Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The data was collected over 2 years in 767 patients, including mental examinations and genotype in order to predict the progression of AD over time. Progression is quantified by the change in MMSE score over a 24 month period (ΔMMSE).

2.2. Approach

We aim to develop an algorithm that is robustly accurate across data sets, by creating an ensemble model of the top models tried previously (simple logistic regression, random forests, and Bayesian nets). By weighting our ensemble with boosting, we will try to create an ensemble model that is superior in accuracy to any of the constituent models.

3. Results

4. Discussion

5. Next Steps

5.1. Integrating other learning algorithms

5.2. Gaining Domain Expertise

Several of the researchers who developed the ADNI database are here at Penn. We are being advised by Dr. Leslie Shaw and Dr. John Trojanowski on the best usage of the database and the relationship between the data features and the disease.

6. Final Report

Your final project report can be at most 5 pages long (include all text, appendices, figures, references, and anything else), and must be written in the provided LaTeX template.

At a minimum your final report must describe the problem/application and motivation, survey related work, discuss your approach, and describe your results/conclusions/impact of your project. It should include enough detail such that someone else can reproduce your approach and results. For inspiration on what should be included, see the project reports available on the links provided in Section **??**. You will likely end up with a better report if you start by writing a 6-7 page report and then edit it down to 5 pages of well-written and concise prose.

In addition, your report must also include a figure that graphically depicts a major component of your project (e.g., your approach and how it relates to the application, etc.). Such a summary figure makes your paper much more accessible by providing a visual counterpart to the text. Developing such a concise and clear figure can actually be quite time-consuming; I often go through around ten versions before I end up with a good final version.

7. Optional Suggestions for Your Paper and Formatting Guidance

7.1. Figures

You may want to include figures in the paper to help readers visualize your approach and your results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, be sure to include a caption describing your figure.

You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX), but always place two-column figures at the top or bottom of the page.

7.2. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

Algorithm 1 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9± 0.2	96.7± 0.2	√
CLEVELAND	83.3± 0.6	80.0± 0.6	×
GLASS2	61.9± 1.4	83.8± 0.7	√
CREDIT	74.8± 0.5	78.3± 0.6	
HORSE	73.3± 0.9	69.7± 1.0	×
META	67.1± 0.6	76.5± 0.5	√
PIMA	75.1± 0.6	73.9± 0.5	
VEHICLE	44.9± 0.6	61.5± 0.4	√

7.3. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table, as in Table 1.

Tables contain textual material that can be typeset, as contrasted with figures, which contain graphical material that must be drawn. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns, but place two-column tables at the top or bottom of the page.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Doerflinger, D. Carolan. How to try this: The mini-cog. *Elektronika IR Elektrotechnika*, 107(12):62–71, 2007.

Duthey, B. Background paper 6.11 alzheimer disease and other dementias. Technical report, World Health Organization, Paris, France, 2013.

Umer, R. *Machine learning approaches for the computer aided diagnosis and prediction of Alzheimer's disease based on clinical data*. PhD thesis, Department of Computer Science, University of Georgia, 2011.