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, with plans to supplement our data with work from additional longitudinal studies. We found disappointing results when using a linear regression to predict disease progression and determined to switch to a classification problem, for which we can apply several more machine learning algorithms. 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)

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)

Table 1. Interpretations of the MMSE Score, ranging from full normal cognition to sever impairment – which is closely correlated with the presence of dementia.

Cognitive States	MMSE Score (out of 30)
Normal Cognition	≥ 27 points
Mild Impairment	19 - 24 points
Moderate Impairment	10 - 18 points
Severe Impairment	≤ 9 points

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.

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

First, we split the data into training and testing sets. As we were interested in creating an algorithm that could predict disease progression quantitatively, we applied a linear regression to predict (ΔMMSE). Since several of our features were categorical data, we used dummy variables to break them into multiple features that could be weighted accordingly in the model.

Next, we developed a system to categorize patients by Normal Cognition, Mild Impairment, Moderate Impairment, and Severe Impairment; based on MMSE scores (Table 1). We used these classes to begin applying classification algorithms, including SVM, Decision Tree, and Naive Bayes.

3. Results

3.1. Linear Regression

With an 80:20 training/testing split, the linear regression predictor explained 60% of the variance in our (ΔMMSE). The residual sum of squares was 10.42.

3.2. Classification Algorithms

4. Next Steps

4.1. Integrating other learning algorithms

Given the initial promising results after switching from a regression problem to a classification problem, we are focused on developing our suite of classification algorithms. We are going to apply ensemble methods on the collection of algorithms we tried.

4.2. Supplementing the dataset

We are going to include data from a study done in 2012, the AddNeuron trial. This includes additional clinical evidence and better genomic features.

4.3. 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.

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.