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 tried five different classifiers with different parameters and reported the best results. 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)

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.

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).

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

We used these classes to begin applying classification algorithms, including SVM, Decision Tree, and Naive Bayes.

Table 1. Interpretations of the MMSE Score, ranging from full normal cognition to sever impairment – which is closely correlated with the presence of dementia. We used this breakdown for classification style ML methods.

Cognitive States	MMSE Score (out of 30)
Normal Cognition	> 27 points
Mild Impairment	27 - 23 points
Moderate Impairment	23 - 20 points
Severe Impairment	< 20 points

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. When trying other regressors, such as support vector regression, we found similar results.

3.2. Classification Algorithms

After switching to a classification problem, we were able to obtain better results. We tried five different classification algorithms, SVM, K-Nearest Neighbors, Naive Bayes, Decision Trees and ADABoost, on different parameters. The SVM model with a penalty of 0.1 and a linear kernel gave us the best results, with a training accuracy of 64.4% and testing accuracy of 67.5%.

Table 2. Classification accuracies for different variations the machine learning models SVM, Decision Trees, and Naive Bayes on dataset categorized by MMSE labels from Table 1.

Model		Params	Train	Test Acc.
SVM	C = 0.1	Linear	64.4%	67.5%
SVM	C = 3	Linear	62.8%	66.2%
SVM	C = 0.5	Gauss	70.0%	65.6%
SVM	C = 1	Poly	70.0%	61.0%
KNN	K = 5		65.9%	65.6%
N. Bayes	Multi	nomial	62.5%	65.6%
D. Trees	Depth	2	61.5%	65.6%
Adaboost	D.Tree		57.1%	59.7%

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. Going forward, we will work with other machine learning classifier combinations as well as looking for extract and compute features from the dataset to feat into the classifiers.

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. Furthermore, we are also considering complementing our database with the Penn AD Core Center database. Namely, we want to look specifically at some of the genome sequencing data to attempt to derive meaning from imputed genomes.

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, especially in relation to data where domain knowledge is critical to meaningful feature extraction.

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.