Relating Biological and Clinical Features of Alzheimer's Patients With Predictive Clustering Trees

Martin Breskvar[1,2]
martin.breskvar@ijs.si

Bernard Ženko[1]
bernard.zenko@ijs.si

Sašo Džeroski[1,2]
saso.dzeroski@ijs.si

[1]Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[2]Jožef Stefan Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

ABSTRACT

This paper presents experiments with Predictive Clustering Trees that uncover several subpopulations of the Alzheimer's disease patients. Our experiments are based on previous research that identified the everyday cognition as one of the most important testing domains in the clinical diagnostic process for the Alzheimer's disease. We are investigating which biological features have a role in the progression of the disease by observing behavioral response of the patients and their study partners. Our dataset includes 342 male and 317 female patients from the ADNI database that are described with 243 clinical and biological attributes. The resulting clusters, described in terms of biological features, show behavioral and gender specific differences between clusters of patients with progressed disease. These findings suggest a possibility that the Alzheimer's disease is manifested through different biological pathways.

1. INTRODUCTION

Alzheimer's disease (AD) is a form of dementia, which represents a large portion of all dementias. It is a neurodegenerative disease affecting many aspects of the patients life, including physical, psychological and social wellbeing. This inevitably leads to severe decrease of life quality. Currently about 47.5 million people worldwide suffer from dementia,[1] and its incidence is expected to triple by the year 2050.

In order to diagnose AD with certainty a histopathologic examination has to be conducted, which is the main reason why in practice AD diagnosis is mainly based on clinical criteria that can be subjective. Finding links between the clinical and biological characteristics of the disease is therefore an important research topic: its advancement could potentially improve the understanding of the disease pathophysiology and enable its detection at earlier stages.

In this work, we address the problem of finding possible connections between biological and clinical features of AD patients with the use of Predictive Clustering Trees (PCTs). Our goal is not to provide a model for diagnosing the disease, but rather to cluster patients into homogeneous groups that share biological features. This way we should be able to investigate the traits of the grouped patients in more detail. One of the most distinctive properties of PCTs is their ability to learn models for predicting structured or complex variables, e.g., vectors, time-series or hierarchies. We use a dataset of Alzheimer's patients obtained from the ADNI database.[2] By using PCTs, we were able to construct clusters homogeneous in respect of several clinical variables simultaneously and not just a single one as with standard decision trees.

The remainder of the paper is structured as follows. Section 2 presents the dataset, methodology and the experimental design. Section 3 describes the results. Finally, in Section 4 we analyze the results and present our conclusions.

2. DATA AND METHODOLOGY

2.1 The Data

All data used comes from Alzheimer's Disease Neuroimaging Initiative (ADNI) database[2]. ADNI is an international observational study of healthy elders, people with mild cognitive impairment (MCI) and people with Alzheimer's disease. It collects a wide range of clinical and biological data for each patient at multiple time points. We used the ADNIMERGE table, which is a joined dataset from multiple ADNI data collection domains.

The dataset includes information on 659 patients (342 male, 317 female). Each patient is described with 56 biological and 187 clinical attributes. Some numerical values have been transformed in order to make them more linear. Out of 243 attributes, 74 contain missing values.

Biological attributes include ABETA peptides, APOE4 genetic variations, intracerebral volume (ICV), results from many laboratory measurements like glucose and protein levels, red and white blood cell counts, MRI volumetric data, (Ventricles, Hippocampus, WholeBrain, Entorhinal gyrus,

[1] Source: World Health Organization (march 2015).

[2] The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations. More information can be found at http://www.adni-info.org and http://adni.loni.usc.edu.

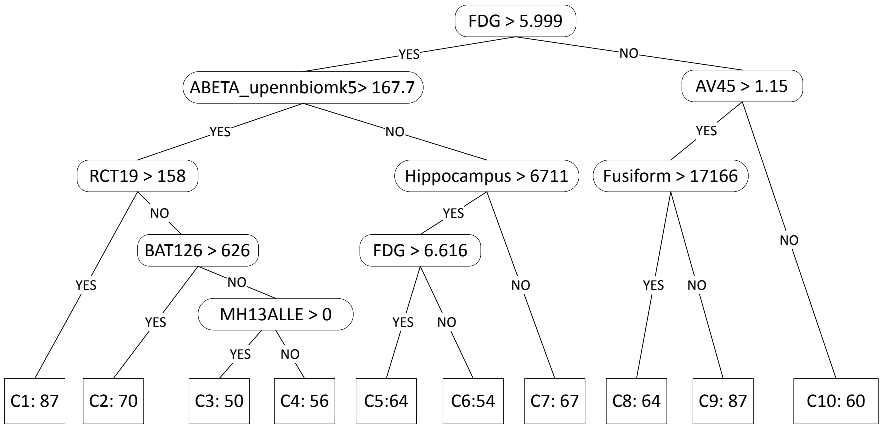


Figure 1: Predictive Clustering Tree, showing 10 clusters (cluster IDs and numbers of patients in each cluster).

Fusiform gyrus, Middle temporal gyrus), TAU and PTAU proteins, and PET imaging results (FDG-PET and AV45).

Clinical attributes include Alzheimer's Disease Assessment Scale (ADAS13), Mini Mental State Examination (MMSE), Rey Auditory Verbal Learning Test (RAVLT), which is divided into several different stages (immediate, learning, forgetting and percantage of forgetting), Functional Assessment Questionnaire (FAQ), Montreal Cognitive Assessment (MOCA) and Everyday Cognition, which consists of questions that are answered by patients themselves (ECogPt) and their study partners[3] (ECogSP). Again, this cognitive evaluation consists of several domains (Memory, Language, Organization, Planning, Visuospatial abilities, Divided attention and Total score). Also Neuropsychiatric Inventory Examination, Neurological Exam, Modified Hachinski Ischemia Scale, Geriatric Depression Scale, Baseline symptoms (nausea, vomiting, diarrhea, sweating, etc.), Clinical Dementia Rating (CDR), Medical History, patient gender and handedness have been included.

The diagnosis (DX) that has been given by the physician at the first examination is also included in the data. The possible values for the DX attribute are Cognitively Normal (CN), Significant Memory Concern (SMC), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI) and Alzheimer's Disease (AD). The diagnosis distribution is the following: CN=173, SMC=94, EMCI=148, LMCI=134, AD=110.

We are using only the baseline data (i.e., data gathered when patients enrolled in the ADNI study and have been examined and tested for the first time).

2.2 Experimental design

In our study we are especially interested in the everyday cognition of patients, therefore we will give a brief overview of the everyday cognition, as it is understood and evaluated within the ADNI database. Everyday cognition (ECog) is a questionnaire, that requires cooperation of both patient and his or her study partner. It assesses the patient's capability to perform normal, everyday tasks. Patients and study partners must individually compare the patient's current activity levels and capabilities with levels from 10 years prior the examination. The domains of memory, language and executive functioning are assessed. Answers are evaluated on a 5 point scale: (1) no change or performing better, (2) occasionally performs worse, (3) consistently performs worse, (4) performs much worse, (5) does not know. According to Farias et. al.[4], everyday cognition shows promise as a tool for measuring general and domain-specific everyday functions in the elderly. We have decided to design our experiment on that assumption and we aim to connect existing biological and clinical features in order to observe differences of predicted values between clusters.

We have used Predictive Clustering Trees for the task of multi-target prediction. Our targets were all the ECog components and the diagnosis itself. Our descriptive space was defined by all the laboratory measurements, neuropathology, medical history and gender. We have included medical history in the descriptive space because we wanted to observe whether pre-existing conditions such as alergies play a role in the disease progression. Additionally we included gender, because according to Barnes et. al.[1] gender specific differences do exist. We have pre-pruned our clustering tree with the constraint of minimum 50 examples per leaf.

2.3 Predictive Clustering Trees

The concept of predictive clustering was introduced in 1998 by H. Blockeel [2] and can be seen as a generalization of supervised and unsupervised learning. Even though predictive modeling and clustering are usually viewed as two separate tasks, they are connected by the methods that partition the instance space into subsets. We can also consider these methods to be clustering methods. An example of such methods are decision trees.

If we consider a decision tree in the predictive clustering paradigm, the tree is a hierarchy of clusters. We refer to those trees as predictive clustering trees (PCTs). An obvious benefit of PCTs is that they, in addition to predictions, also provide symbolic descriptions of the clusters.

[3] Each patient must have a study partner, a person who is in frequent contact with the patient, provides information about the patient and is able to independently evaluate the patient's functioning.

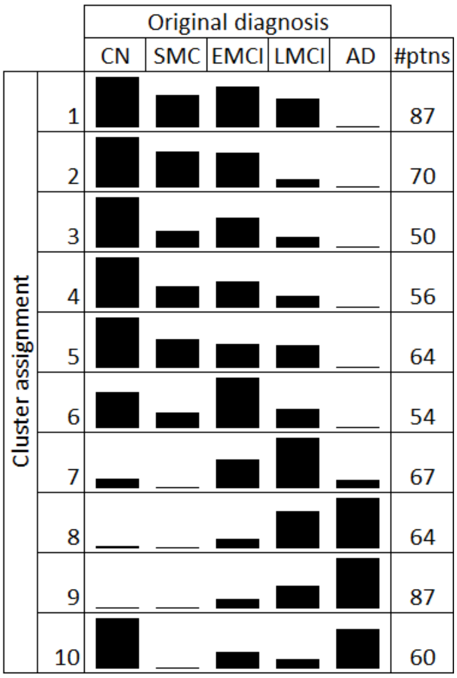


Figure 2: Normalized distribution of original diagnoses with respect to the clusters modeled by the PCT in Figure 1.

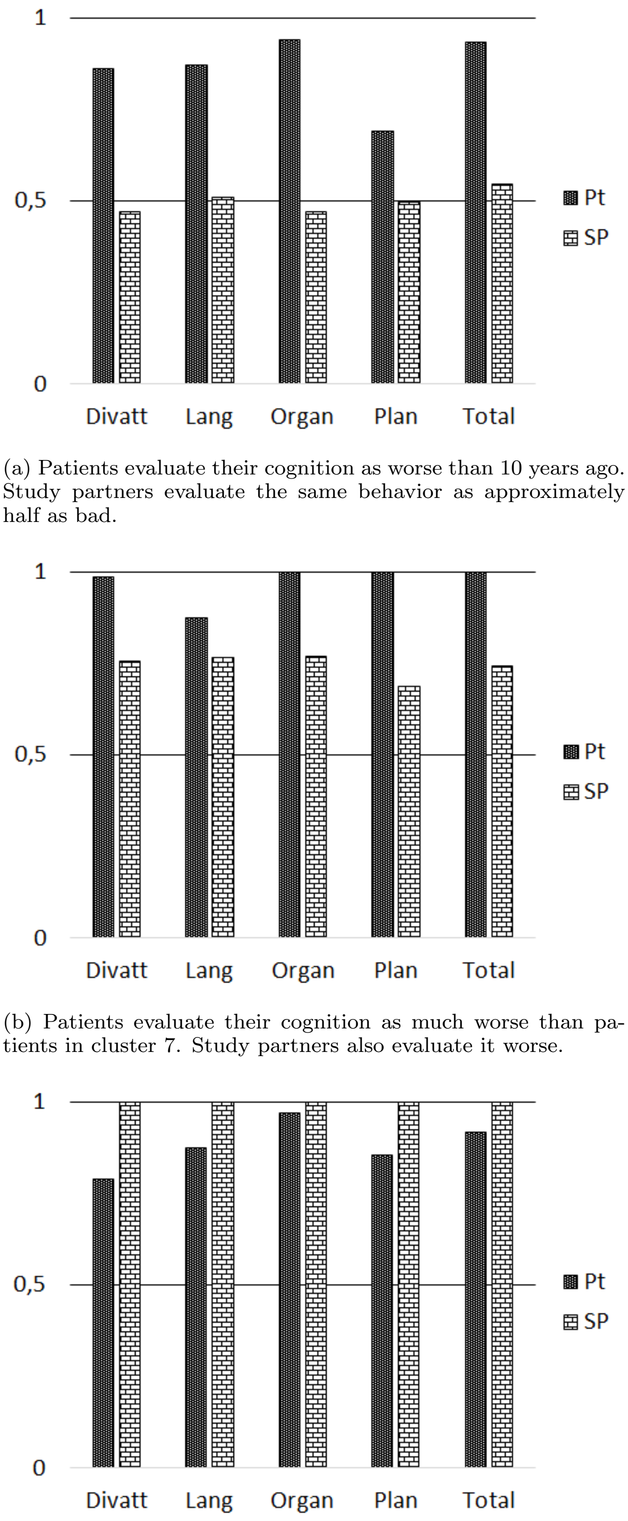
Each node in the clustering tree represents a cluster and has a symbolic description (except for the root node) in the form of a conjunction of conditions on the path from the root node to the selected cluster node. In case of the PCT in Figure 1, the examples in the root node are split according to condition $FDG > 5.999$. Examples, whose value of the FDG attribute is greater than the value 5.999 will go to the left branch, the others to the right branch. On the next level of the clustering tree, nodes $AV45 > 1.15$ and $ABETA\_upennbiomk5 > 167.7$ are now split again iteratively until we reach leaf nodes $C1$. Examples in cluster $C1$, for example, are those that correspond to the condition: $FDG > 5.999$ & $ABETA\_upennbiomk5 > 167.7$ & $RCT19 > 158$.

PCTs support multi-target predictions which means we can learn a model with respect to not only one target variable but many. This gives us the tool needed to predict complex structures that can also be interconnected.

Several different predictive clustering methods [3, 5, 6, 7, 8, 9] are implemented in the software package CLUS (available at http://sourceforge.net/projects/clus/).

3. RESULTS

The result of our analysis is the PCT presented in Figure 1. We have investigated all ten clusters in the leaf nodes and Figure 2 shows relative distribution of original diagnoses (DX) in all the clusters. Clusters 1 to 6 are relatively diverse and we can state that the presence of Alzheimer's patients in these clusters is unlikely. With the exception of cluster 6, cognitively normal patients are dominant. Cluster 6 also contains patients in the early stage of the disease (EMCI) as well as some LMCI patients. We have focused our atten-

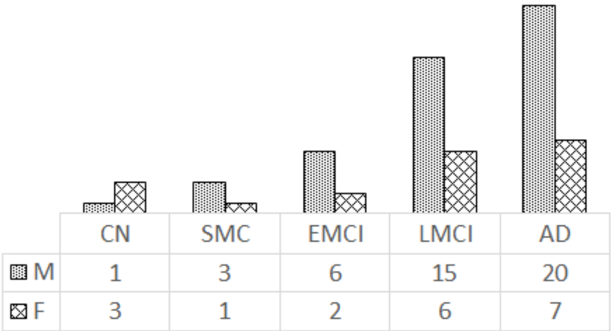


(a) Patients evaluate their cognition as worse than 10 years ago. Study partners evaluate the same behavior as approximately half as bad.

(b) Patients evaluate their cognition as much worse than patients in cluster 7. Study partners also evaluate it worse.

(c) Patients evaluate their behavior milder as their study partners.

Figure 3: Normalized Everyday cognition (ECog) predictions for clusters 7 (3a), 8 (3b) and 9 (3c).



	CN	SMC	EMCI	LMCI	AD
▩ M	1	3	6	15	20
☒ F	3	1	2	6	7

Figure 4: Gender difference in cluster 8.

tion on clusters 7, 8 and 9 because they mainly consist of patients in the stages of late MCI or already developed AD. We have examined the profiles of predicted ECog features. The normalized predictions are shown in Figure 3. Cluster 10 is interesting in the sense that it includes two extremes, healthy patients and heavily affected patients. We assume that this cluster should be further split into two more homogeneous clusters. The exploration of this cluster is planned for further work.

Patients in cluster 7 (Fig.3a) evaluate their cognition as worse than 10 years ago. Their study partners evaluate the same behavior as approximately half as bad. The majority of patients have early and late MCI and the predictions for this cluster correspond to the distribution in Figure 2 quite well. In cluster 8 (Fig.3b), where the majority classes are AD and LMCI, patients evaluate their behavior worse than those in cluster 7. Study partners in this cluster see the situation worse than study partners in cluster 7. In both clusters 7 and 8 the patients always evaluate their behavior worse than their study partners.

In cluster 9 we observe a change in this perception. Study partners evaluate the patients' behavior worse than the patients themselves. This observation could be a direct result of the disease progression, since cluster 9 consists mainly of heavily affected AD patients. On the other hand it could indicate a new disease signature.

We have also analyzed the gender distribution within clusters 7, 8 and 9. We discovered that cluster 7 is gender balanced. Cluster 8 contains more male patients and this dominance is exhibited for all diagnoses as shown in Figure 4. Cluster 9 on the other side contains more women. Specifically, differences occur in classes EMCI and LMCI.

In addition to identifying a cluster of severely affected males and establishing a difference of perception between the patients and their study partners, we have also identified some important features that show potential for discovering specialized clusters. Our results show that AV45, FDG, hippocampal and fusiform volumes and ABETA_upennbiomk5 play an important role in the description of our clusters. As we already mentioned in Section 2.2, we have pre-pruned our clustering tree. The unpruned tree reveals additional important features such as the volume of entorhinal cortex, several laboratory measurements, including glucose level, PTAU_upennbiomk5, and white blood cell count.

4. CONCLUSIONS

This work presents an application of predictive clustering trees to the problem of discovering connections between biological and clinical features of patients with Alzheimer's disease. The result is a PCT with ten clusters, three of which are interesting. We have analyzed all three and discovered interesting indications that biological features have an impact on the observed clinical behavior of the patients.

We have also discovered a gender specific differences, as we have initially expected in the design of the experiment. We have identified several biological features that might be connected with the Alzheimer's disease progression. The results are promising and in line with other studies, but additional research will need to be conducted in order to further validate the current results presented here.

5. ACKNOWLEDGMENTS

We would like to acknowledge the support of the Slovenian Research Agency (through a young researcher grant to MB and the programme grant Knowledge Technologies) and the European Commission (through the projects MAESTRA – Learning from Massive, Incompletely annotated, and Structured Data - grant FP7-ICT-612944) and HBP – The Human Brain Project - grant FP7-ICT-604102).

6. REFERENCES

[1] L. L. Barnes, R. S. Wilson, J. L. Bienias, J. A. Schneider, D. A. Evans, and D. A. Bennett. Sex differences in the clinical manifestations of alzheimer disease pathology. *Archives of General Psychiatry*, 62(6):685–691, 2005.

[2] H. Blockeel. *Top-Down Induction of Clustering Trees.* PhD thesis, Katholieke Universiteit Leuven, Department of Computer Science, 1998.

[3] H. Blockeel and J. Struyf. Efficient algorithms for decision tree cross-validation. *The Journal of Machine Learning Research*, 3:621–650, 2003.

[4] S. T. Farias, D. Mungas, B. R. Reed, D. Cahn-Weiner, W. Jagust, K. Baynes, and C. DeCarli. The measurement of everyday cognition (ecog): scale development and psychometric properties. *Neuropsychology*, 22(4):531, 2008.

[5] D. Kocev, C. Vens, J. Struyf, and S. Džeroski. Ensembles of multi-objective decision trees. *Machine Learning: ECML 2007*, pages 624–631, 2007.

[6] I. Slavkov, V. Gjorgjioski, J. Struyf, and S. Džeroski. Finding explained groups of time-course gene expression profiles with predictive clustering trees. *Molecular BioSystems*, 6(4):729–740, 2010.

[7] J. Struyf and S. Džeroski. *Constraint based induction of multi-objective regression trees.* Springer, 2006.

[8] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008.

[9] B. Ženko. *Learning Predictive Clustering Rules.* PhD thesis, University of Ljubljana, Faculty of Computer and Information Science, 2007.