

Classification of dementia types from cognitive profiles data

Giorgio Corani¹, Chris Edgar², Isabelle Marshall², Keith Wesnes², and Marco Zaffalon¹

¹ IDSIA (Istituto Dalle Molle di Studi sull'Intelligenza Artificiale)
Manno, Switzerland

`giorgio[zaffalon]@idsia.ch`

² Cognitive Drug Research Ltd

Goring-On-Thames, U.K.

`chrise[isabellem,keithw]@cognitivedrugresearch.com`

Abstract. The Cognitive Drug Research (CDR) system is specifically validated for dementia assessment; it consists of a series of computerized tests, which assess the cognitive faculties of the patient to derive a *cognitive profile*. We use six different classification algorithms to classify clinically diagnosed diseases from their cognitive profiles. Good accuracy was obtained in separating patients affected by Parkinson's disease from demented patients, and in discriminating between Alzheimer's disease and Vascular Dementia. However, in discriminating between Parkinson disease with dementia (PDD) and dementia with Lewy bodies (DLB), the accuracy was only slightly superior to chance; the existence of a significant difference in the cognitive profiles of DLB and PDD is indeed questioned in the medical literature.

Keywords: CDR computerized assessment system, dementia, classification

1 Introduction

Dementia is one of the most common disorders among the elderly; it causes a progressive decline in cognitive functions such as memory, attention and language. The Cognitive Drug Research (CDR) system [1] is widely used in clinical trials and has been specifically validated for use in dementia; it consists of a series of computerized tests (*tasks*), which assess some cognitive faculties of the patient, such as memory, attention, reaction times. The set of the measures collected during all tasks represents the *cognitive profile* of the patient.

In [3] the cognitive profiles returned by the CDR test are used to address two different classification problems: (a) to discriminate between demented patients and controls, and (b) to discriminate from among the different types of dementia. An accuracy higher than 90% was obtained on both tasks by using the Naive Credal Classifier, a generalization of the Naive Bayes Classifier to imprecise probabilities, or credal sets.

In this paper, we propose a similar approach, but a few important differences in the data: (i) the number of tasks selected from the CDR battery is smaller, and mainly restricted to the attentional measures. Indeed, it is not clearly stated in literature whether the attentional measures of the cognitive profile are really different between some kinds of dementia, and therefore the subject is worthy of investigation. Moreover, a visit including only attentional tasks would take no more than 10 mins., while a complete administration of the CDR tasks would take between 30 and 45 mins. (ii) Standard deviation and number of outliers of each featured measure are available, while only the median was available in the previous study; indeed, fluctuations of cognitive faculties, captured by the standard deviation of the variables, are important to characterize the cognitive profiles, as shown in [4]; (iii) an enlarged set of dementias is considered, and Parkinson’s patients are used instead of healthy controls to assess whether the system is able to discriminate between motor impairment and dementia. (iv) A peculiar investigation of this study is moreover the inter-comparison of the accuracies obtained using the cognitive profiles assessed at the first visit and at the third visit on the CDR test. Indeed, although patients are usually trained twice in the clinical practice on the tests prior to the definitive assessment of the cognitive profile at the third visit, performing the classification directly on the first-visit data could allow for time and money savings. We experimentally check whether first-visit data leads to classification accuracy better or statistically not different from third-visit data (hypothesis H_0). If H_0 is verified, we can indeed easily recommend the use of first-visit data. If on the contrary the experiment show a statistical improvement of the accuracy using third-visit data (hypothesis H_1), the judgment becomes more complex.

The aim of this paper is to better understand the domain of dementia analysis through cognitive profiles, which has been only rarely explored up to now by means of ML techniques, and to provide findings useful to ML scientists that will in the future work on similar data. We have looked for *robust* experimental findings, supported by the results of a set of classification approaches, rather than fine-tuning for performance a specific algorithm. We have therefore considered a set of different classification algorithms, including very well-known approaches such as J4.8, naive Bayes, logistic regression and other algorithms which performed well on our data (classification via regression, lazy, etc.). Further classification approaches have been excluded from the analysis because their performance was remarkably worse compared to that of the algorithms eventually selected.

2 The CDR test

Currently, there are more than 30 studies referenced in dementia literature based on the CDR system. The system is entirely computerized, thus allowing for precise measurement of the latency of each response.

The following tasks are considered in this study:

- *Simple reaction time (SRT)*: the patient should press the “yes” button as quickly as possible as the word “yes” is displayed on the monitor. The task is repeated 30 times.
- *Digit VIGilance task (VIG)*: a random target digit is constantly displayed on the monitor screen. A series of digits are then presented and the patient should press “yes” as quickly as possible as the digit in the series matches the target digit.
- *Choice reaction time (CRT)*: either the word “no” or the word “yes” is displayed and the patient should press the corresponding button as quickly as possible. 30 trials are performed.
- *Delayed PICTure recognition (DPIC)*: a series of 14 pictures is presented on the monitor for the patient to remember. Afterwards, the same pictures are presented to the patient, together with 14 distracting pictures; for each picture the patient had to indicate whether or not it belongs to the first series.

For each task, several index of performances are recorded.

3 The dataset

Two separate datasets contain the cognitive profiles assessed at the first and at the third visit on the CDR tasks. The dataset of the first visits contained 1842 records of cognitive profiles, while the dataset of the third visit contained 1670 records. Data were taken from patients before they entered different clinical trials. The datasets used in this study contains patients from Western Europe, Eastern Europe and Asia.

Different kinds of dementia are present in the dataset: Alzheimer disease (AD), Dementia with Lewy Bodies (DLB), Parkinson Disease with Dementia (PDD), Vascular Dementia (VAD); moreover, the dataset comprised patients affected by Parkinson’s Disease (PD). PD patients are actually *not* demented, though they suffer a significant motor impairment. The distribution of the diseases is as follows: {AD 21%; DLB 10%; PDD 28.5%;VAD 34.5%; PD 6%} and it is almost identical for the datasets of the first and of the third visit.

In particular, there is not yet gold standard for the clinical distinction of AD and VAD, although there is the need of distinguishing between them because of the differences in the necessary treatments.

PDD and DLB are also very similar diseases: they are characterized by both dementia and parkinsonism, and their cognitive profiles have been found to be striking similar [4].

Therefore, three classification tasks appear as of scientific interest:

- *task 1*: to classify patients into three macro-classes as (AD-VAD, PDD-DLB, PD);
- *task 2a*: to discriminate between PDD and DLB;
- *task 2b*: to discriminate between AD or VAD. This is the only task for which features related to Delayed PICTure Recognition are available.

The tasks have been implemented separately from each other, and evaluated separately in this paper. However, in a real clinical use, they could be implemented within an actual cascade classification system, the output of classifier 1 possibly feeding, depending on the classification output, classifier 2a or 2b. Such an architecture would allow to easily use the additive DPIC features for tasks 2b.

4 Experimental Setting

Datasets have been analyzed using six different classification algorithms, implemented within the open source³ WEKA software [5]. Classification algorithms have been used with their default settings.

We relied on the indication of the CDR staff as for the set of features to be used. All the considered features are numerical; however, we discretized them through the MDL-based supervised discretization algorithm originally proposed in [6]. Indeed, the experimental investigation carried out in [7] showed that quite frequently the use of such discretization algorithm leads to improved accuracy compared to the raw data. This turned out to be the case even for our dataset; for instance, J4.8 improved of about 2 accuracy points thanks to the use of discretized data.

The accuracy of the classifiers has been assessed via 10 runs of 10-fold cross validation. The statistical significances of the differences in accuracy have been tested via a *t*-test (5% significance); in particular, to properly manage the cross-validation errors, we used the corrected resampled *t*-test implemented in WEKA.

For tasks (1) and (2a) the features related to simple reaction time, choice reaction time, digit vigilance have been used; for task (2b), also the features related to picture recognition have been used.

5 Results

5.1 Classification task 1: {(AD/VAD), (PDD/DLB), (PD)}

The accuracies obtained on this task are shown in Table 1. Depending on the classification algorithm, the classification accuracy ranges between 75% and 80%.

None of the 7 classification approaches showed a significant difference of accuracy working on the data of the first visit rather than on the data of the third visit; therefore, these results clearly support H_0 .

Classification-via-regression and logistic regression appear as the best performing approaches. However, the objective of this study was mainly to get an indication about the obtainable accuracy, rather than fine tuning the algorithms for the best performance.

The confusion matrix for Logistic Regression is reported in Table 2; it shows that most misclassifications occurs between (AD-VAD) and (PDD-DLB), while it

³ Available at the URL: <http://www.cs.waikato.ac.nz/~ml/index.html>.

classifier	1st visit		3d visit		Significant difference?
	average	std. dev.	average	std. dev.	
<i>NAVE BAYES</i>	75.81	3.18	74.90	3.46	NO
<i>BAYES NETWORK</i>	75.74	3.13	74.82	3.40	NO
<i>J4.8 TREE</i>	78.46	3.06	77.49	3.04	NO
<i>SMO</i>	78.12	2.89	78.11	2.90	NO
<i>LOGISTIC REGRESSION</i>	80.09	3.09	78.82	2.89	NO
<i>LAZY.LBR</i>	78.30	2.94	77.17	3.40	NO
<i>CLASS. VIA REGR. (M5)</i>	79.93	2.81	78.51	2.98	NO

Table 1. Accuracy of different classification algorithms in task 1.

	<i>AD-VAD</i>	<i>DLB-PDD</i>	<i>PD ← classified as:</i>
<i>AD-VAD</i>	898	117	6
<i>DLB-PDD</i>	181	510	13
<i>PD</i>	20	30	67

Table 2. Confusion matrix for Logistic Regression on task 1.

is quite rare for Parkinson’s disease to be confused with dementia. In particular, a demented patient was likely to be diagnosed as PD with almost negligible probability, while with slightly higher probability a PD patient is diagnosed as demented. However, this was probably due also to the very low proportion of PD in our dataset (6%).

5.2 Classification task 2a: {(PDD), (DLB)}

classifier	1st visit		3d visit		Significant difference?
	average	std. dev.	average	std. dev.	
<i>NAVE BAYES</i>	56.00	6.51	58.29	7.03	NO
<i>BAYES NETWORK</i>	56.08	6.34	58.32	7.02	NO
<i>J4.8 TREE</i>	54.73	6.24	57.68	7.28	NO
<i>SMO</i>	55.63	6.19	57.23	6.90	NO
<i>LOGISTIC REGRESSION</i>	55.91	6.50	58.00	7.26	NO
<i>LAZY.LBR</i>	55.92	6.65	58.29	7.03	NO
<i>CLASS. VIA REGR. (M5)</i>	55.77	6.64	56.58	7.11	NO

Table 3. Accuracy of different classification algorithms in discriminating between PDD and DLB on a *balanced* dataset.

By running the classifier on datasets containing all the instances of PDD and DLB patients for first and third visit, we measured an accuracy between 70% and 76%.

However, considering that (a) the ratio between PDD and DLB patients in the dataset was about 3:1, and (b) that the similarity of the cognitive profiles of PDD and DLB patients has been reported to be striking [4], we suspected that the classifiers learned to predict the majority class, rather than effectively discriminating between the two classes.

To check our hypothesis, we built *balanced* datasets, containing the same number of PDD and DLB patients. The first-visit dataset contained 178+178 patients, and the third-visit dataset 155+155 patients. The results are reported in Table 1.

The accuracy ranged between 54% and 58%; it was just slightly superior to the 50% of a random guess. Also in this case, no significant differences were found using either the first-visit or the third-visit data. However, the most important finding of our analysis is that it is not possible, *regardless the used data*, to reliably discriminate between the two diseases starting from the cognitive profiles. The cognitive profiles of the two diseases were so similar that 9 out of the 11 features of the cognitive profile were discretized into a unique bin, i.e. they were useless to discriminate between PDD and DLB.

We think these results to be mainly due to the largely overlapping features of the two diseases; indeed, a recent clinical paper [8] states in its conclusions: “*from the pathologist’s point of view, the brains of PDD and DLB patients do not present reliably distinctive features. Therefore, it is probable that in the near future PDD and DLB will be recognized as the same disease with two different courses*”.

On the base of these findings, it seems hence advisable to merge these two diseases into a unique class in future classification works.

5.3 Classification task 2b {(VAD), (AD)}

classifier	1st visit		3d visit		Significant difference?
	average	std. dev.	average	std. dev.	
NAVE BAYES	68.82	5.48	68.38	5.38	NO
BAYES NETWORK	68.27	5.44	68.34	5.37	NO
J4.8 TREE	70.08	4.92	68.96	5.38	NO
SMO	70.14	5.05	69.45	5.16	NO
LOGISTIC REGRESSION	70.18	4.98	69.19	5.11	NO
LAZY.LBR	69.14	5.21	68.77	5.26	NO
CLASS. VIA REGR. (M5)	69.70	4.87	68.58	5.10	NO

Table 4. Accuracy of different classification algorithms in discriminating between VAD and AD on a *balanced* dataset.

By running the classifier on datasets containing all the instances of VAD and AD patients for first and third visit, we measured an accuracy between 71% and 74%. Also in this case, none of the classifiers showed a significant difference in accuracy working on first-visit or third-visit data.

The difference between the accuracy recorded on first-visit and third-visit was lower than 2 points of accuracy for each classifier; nevertheless, 6 classifier out of 7 showed a better accuracy on the data of the first visit (2 times such improvement was found to be significant). Overall, all classifiers support hence H_0 .

However also in this case, we wanted to avoid the prevalence of VAD patients in the dataset (VAD/AD about 1.6 in the dataset), to bias the experimental results; therefore, we cross-checked these results by running the classifiers also on balanced datasets.

The balanced dataset of the first visit contains 355+355 patients, and that of the third visit 346+346 patients. The accuracy is in this case around 70%, with narrow differences between the different algorithms, as reported in Table 4. Even the differences between the accuracies measured on the first visit data and the third visit data are narrow; although 6 classifiers out of 7 show a slightly higher accuracy on the first-visit data, in no case such difference is statistically significant. Indeed, also these results fully support H_0 .

We finally report that a few features (mainly related to simple reaction time and choice reaction time) were discretized into a single bin, being in practice useless for classification.

6 Conclusions

The medical literature acknowledges the need for further research to improve clinical definitions of dementia and to determine the utility of various standardised instruments in increasing diagnostic accuracy, which currently average 58% in DLB, 50% in VAD and 81% in AD. The ability to apply a single diagnostic assessment to a range of dementias, with good average sensitivity equivalent to, or above that seen with current assessments, is hence a useful addition to existing assessment tools and diagnostic criteria.

This paper provides some clear conclusions regarding the analysis of cognitive profiles for dementia screening via ML techniques; in particular, we (a) looked for robust results, inter-comparing the findings obtained by using different classification algorithms and (b) we integrated our empirical results with the domain-specific literature. Both such approaches are recommended when a previously unexplored domain has to be investigated via ML techniques.

Our experimental results show that with accuracy up to 80% it is possible to discriminate between Parkinson's disease and the two dementia macro-classes (PDD-DLB and AD-VAD). To reliably discriminate between PDD and DLB starting from cognitive profiles is however not achievable; indeed, the actual

existence of a significant difference between the two diseases is currently strongly debated within the medical literature. On the basis of these findings, we advice to merge these two classes into a unique class in future classification works. A further classification task of interest is to discriminate between AD and VAD dementia; we show that in this case an accuracy up to 70% can be reached. We have moreover found that a number of variables of the cognitive profile is not useful in discriminating between AD and VAD; hence, machine learning algorithms appear to be useful also because they show the variables which are sensitive for classification.

A interesting finding that using the first-visit data instead of third-visit data does not lead to any worsening of the classification accuracy. Thus, we can indeed strongly support the use of first-visit data, thus allowing for time and money savings, from both the patients and the company viewpoint.

Acknowledgments

Marco Zaffalon gratefully acknowledges partial support by the Swiss NSF grant 200020-109295/1.

CDR acknowledges the Department of Trade and Industry for partially funding this research through a Global Watch International Secondment grant.

References

1. Simpson, P., Surmon, D., Wesnes, K., Wilcock, G.: The cognitive drug research computerised assessment system for demented patients: A validation study. *Int. J. of Geriatric Psychiatry* **6** (1991) 95–102
2. Knopman, D.S., DeKosky, S.T., Cummings, J.L., Chui H., Corey-Bloom, J., Relkin, G.: Practice parameter: Diagnosis of dementia (an evidence-based review): Report of the quality standards subcommittee of the American academy of neurology. *Neurology* **56** (2001) 1143–1153
3. Zaffalon, M., Wesnes, K., Petrini, O.: Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data. *Artificial Intelligence in Medicine* **29**(1-2) (2003) 61–79
4. Ballard, C., Aarsland, D., McKeith, I., O'Brien, J., Gray, A., Cormack, F., Burn, D., Cassidy, T. and Starfeldt, R., Larsen, J., Brown, R., Tove, M.: Fluctuations in attention - PD dementia vs DLB with parkinsonism. *Neurology* **59** (2002) 1714–1720
5. Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc,US (2005)
6. Fayyad, U.M., Irani, K.B.: On the handling of continuous-valued attributes in decision tree generation. *Machine Learning* **8** (1992) 87–102
7. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. In: *International Conference on Machine Learning*. (1995) 194–202
8. Sellal, F.: Parkinson's disease with dementia and dementia with Lewy body disease: two syndromes, the same disease? *Psychogeriatrics* **6** (2006) 30–34