Improving Dementia Screening Tests with Machine Learning Methods

William Rodman Shankle
Department of Neurobiology
Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92715-4285

Piew Datta, Michael Dillencourt, Michael Pazzani
Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92715-3425

Correspondence information:
William Rodman Shankle
Department of Neurobiology
University of California, Irvine
Irvine, Ca 92715-4285
Phone number: (714) 824-2382
Fax: 824-3049
Email: rshankle@uci.edu

Abstract

We applied Machine Learning (**ML**) methods to test whether their application on data derived from two simple tests of cognitive and functional skills can improve dementia screening. DSM-IV criteria for dementia were used to categorize patients into normal, cognitively impaired but not demented, and demented groups. The two tests are the *Functional Activities Questionnaire* (**FAQ**) and the *Six-Item Blessed, Orientation, Memory and Concentration Exam* (**BOMC**); both are recommended for screening by the *Agency For Health Care Policy Research* (**AHCPR**).

The sample consisted of the initial visits of 609 normal, cognitively impaired or demented subjects evaluated at the UC Irvine Alzheimer's Disease Research Center (**ADRC**). Using published cutoff criteria for the FAQ (> 8 is demented) and the BOMC (> 10 is demented) tests gave classification accuracies of 73% and 64% respectively. Combining the cutoff criteria gave a 60% classification accuracy.

On this same sample, we applied four ML algorithms using the FAQ and BOMC test response results. ML methods obtained 86% overall classification accuracy, an increase of 13-24% over conventional cutoff criteria. For the cognitively impaired group, classification accuracy was 24% using cutoff criteria, and 61-64% using ML methods, representing a 37-40% improvement.

ML methods extract more information from these AHCPR-recommended tests to increase substantially their accuracy in detecting cognitive impairment and dementia, and to implement their combined use in a simple, descriptive way.

Keywords: Dementia, Cognitive Impairment, Rule-based systems, Expert systems, Machine Learning, Specificity, Sensitivity, Functional activities, BOMC, Classification

Introduction

Although there are a number of medical areas to which Machine Learning (**ML**) systems have been applied, dementia has not been one of them. Yet dementia is a complex problem with correct diagnosis requiring historical data, physical exam, cognitive testing, laboratory studies and imaging. Although diagnostic accuracy using clinical NINCDS-ADRDA criteria for probable AD[1] is about 88% compared to post-mortem diagnosis[2], most demented patients are seen by community physicians who often do not detect dementia[3] or misdiagnosis it[4] . This problem is compounded by the average delay of 4 years between symptom onset and first physician contact[5], which usually relates to the patient's social embarrassment about having a memory problem. At this point of the disease, physicians are less able to slow the progression and minimize debilitating behavioral effects of the dementia. As an example of an intervention which could have greater value if started earlier in the disease, Lubeck et al.[6], reported a 17% reduction in the $200,000 cost of AD patient care using central cholinergic agonists (Tacrine). A simple, unobtrusive method for detecting dementia early in the disease's course would help get patients to seek early evaluation and treatment. Early detection and correct diagnosis can lead to disease-retarding therapies which can slow disease progression and reduce patient and caregiver stress and morbidities.

With the Agency for Health Care Policy Research's[7] recent development of clinical practice guidelines for the assessment and recognition of Alzheimer's disease and related disorders, simple tests have been identified and recommended from a meta-analysis of the dementia literature. After excluding delirium and depression, two forms, the Functional Activities Questionnaire (**FAQ**[8], and the six-item Blessed Orientation, Memory and Concentration test (**BOMC**[9]) can be used to determine if a person has a dementia, which is defined as multiple cognitive impairments with loss of functional skills related to those cognitive impairments without an altered level of consciousness. Although the AHCPR provided a flow chart to assist dementia detection, they left to the clinician how to combine the test results in determining if a patient has a dementia. ML methods can simplify the task of interpreting test results by constructing a set of criteria to classify the patient. In this article, we report the use of the FAQ and BOMC tests in conjunction with four ML systems on a sample of 609 patients and controls evaluated at the UC Irvine Alzheimer's Disease Research Center. The goal was to determine whether ML methods improve the accuracy of these dementia screening tests for classifying a subject as either unimpaired, cognitively impaired, or demented.

Methods

Sample Description

The sample consisted of the initial visits of 609 subjects seen either as controls or as patients. Patients were concerned about their memory at the University of California, Irvine Alzheimer's Disease Research Center (ADRC). Patients received a complete diagnostic eval-

uation consisting of patient and caregiver interviews, general physical and neurological exam, two hours of cognitive testing including the CERAD[10] neuropsychological battery and other selected tests, routine laboratory testing for memory loss, magnetic resonance neuroimaging with or without single photon emission with computed tomography. Controls were either community volunteers or unaffected spouses of patients, and received an abbreviated, 45 minute version of the patient cognitive battery, which consisted of the CERAD plus measures of activities of daily living. They did not receive a medical exam, laboratory testing or neuroimaging unless cognitive or functional testing suggested an impairment. Subjects with delirium were excluded from the analysis.

Classification of Dementia Status

The diagnosis of dementia status, using DSM-IV criteria[11], was based on a review of all the data by the neurologist and neuropsychologist during their diagnostic review session. Each subject was categorized as either unimpaired, or cognitively impaired but not meeting criteria for dementia, or demented. A classification of **dementia** required the presence of multiple cognitive impairments plus functional impairments resulting from the cognitive impairments in the absence of delirium or other non-organic etiologies such as major depression. **Table 1** shows the sample's characteristics.

BOMC and FAQ tests

The BOMC consists of six questions and the FAQ consists of ten questions. Together they provide a simple means of assessing cognitive and functional status. The answers to these questions were extracted from the UCI ADRC relational database of over 2,000 variables per subject visit to compute the BOMC and FAQ scores.

Machine Learning Systems

We applied four ML systems to the sample, namely *C4.5 and C4.5 Rules*[12], *Naive Bayes*[13], *and IB1*[14] from the machine learning library, MLC++[15]. *C4.5* learns a decision tree to classify dementia state based on the input data's values, which are called **attribute values**. C4.5 applies a hill climbing heuristic search[16] to find attributes that split the data set. At each decision point, an information gain metric using entropy[17] is applied to find the best attribute for splitting the cases into homogeneous groups according to their class. This recursive process ends when all or many of the examples in the group are of the same class. During classification, the test case is processed through the tree until a leaf node is reached. The class of the leaf node is the predicted class of the test case. *C4.5 Rules* simplifies *C4.5's* decision tree and converts it into a set of if-then rules.

The *Naive Bayes* ML uses Bayes theorem to calculate, for an unclassified case (E), the probability of each dementia class given the unclassified example's attribute values using the following formula:

$$P(C_i|E) = P(C_i|A_1 = V_{1j} \& ... \& A_n = V_{nj})$$

where C_i is a class, A_i is the ith attribute, V_{ij} is the jth value for A_i, and there are n attributes. If the attribute values are independent this probability reduces to

$$P(C_i|E) = P(C_i) \prod P(A_k = V_{kj}|C_i)$$

where $P(A_k = V_{kj}|C_i)$ and $P(C_i)$ are estimated from the training data. The class with the highest probability is chosen as the class of the test case.

The last ML method applied, *IB1*, is an instance-based learner that stores all training cases during its learning phase. During classification, the test case is compared to all training cases and the distance to each case is computed. For each test case, the distance between its attribute values and those of each training case is computed. IB1 assigns to the test case the class of the training case having the shortest distance to the test case.

Analytical Design

To establish benchmarks, we determined DSM-IV dementia status classification accuracy using the conventional cutoff criteria reported for the FAQ (>8 indicates functional impairment[8]) and BOMC (>10 is demented[7]). Since no criteria are specified for interpreting the combined results of the FAQ and BOMC scores, we stipulated that both tests must fall below the cutoff criteria to classify a subject as unimpaired, or both tests must be at or above the cutoff criteria to classify a subject as demented. Those subjects falling in between these two ranges were classified as cognitively impaired but not meeting criteria for dementia. These results appear in **Table 2**.

To test the potential predictive power of ML systems for classifying dementia status in this clinical sample, we used the normalized total scores of the BOMC and FAQ tests together with their item scores plus patient age, education, sex. These data constitute the *attributes* of the examples used by the ML system. The attributes of a training sample of patient examples are used to generate criteria to classify a patient's DSM-IV dementia state (*normal, cognitively impaired or demented*).

We ran these ML systems with a *training* and a *testing sample*. The ML system learns a description of the dementia classes from the *training sample*, which is randomly selected from the 609 examples. From the attributes, each ML method creates classification criteria to be evaluated with the *testing sample*, which is composed of all examples not included in the training sample. The *classification accuracy* of an algorithm is the number of correct predictions divided by the total number of predictions. To view the effect of increasing the *size of the training sample*, we experimented with training set sizes of 112, 223, 335, and 447 examples by randomly allocating the appropriate numbers of subjects for each training set, and applying each of the four ML algorithms to the FAQ and BOMC data to generate classification criteria, which were then run on the testing sample to compute dementia status classification accuracy. We repeated this process 30 times for each training

sample size to obtain estimates of the variances of the classification accuracies obtained by the four ML algorithms.

Results

Table 1 shows the distribution of subjects by dementia status. As is true of many clinical samples, normal and cognitively impaired subjects represent a smaller fraction of the sample than the demented subjects (6.9% and 17.6% vs. 75.5%). There are more females than males (58.9% vs. 41.1%) with similar distributions across each dementia group (6.7% vs. 7.2% for normals, 13.9% vs. 22.8% for cognitively impaired, and 79.4% vs. 70% for demented). The normal and cognitively impaired groups have fewer subjects over age 65 than the demented group (4.9% and 13.7% vs. 81.4%), fewer subjects with skilled jobs than the demented group (14.3% and 23.3% vs. 62.4%), and fewer subjects with college education than the demented group (10.4% and 24.2% vs. 65.3%).

Table 2 shows the sensitivities, specificities and predictive values obtained using the AHCPR-recommended cutoff scores for the FAQ and BOMC tests without the ML methods. For the FAQ, sensitivity in *detecting dementia* is 82.8% and specificity in *detecting normal* is 100%. However, for the high risk group, *cognitive impairment without dementia*, only 18.7% of these subjects would be classified as having a problem (i.e., dementia). For the BOMC, sensitivity in *detecting dementia* is 74% and specificity in *detecting normal* is 100%. However, for the high risk group, cognitive impairment without dementia, only 7.5% of the cognitively impaired subjects would be classified as having a problem (i.e., dementia). Using the cutoff criteria of the FAQ and BOMC tests combined, one can classify those below the cutoff for both tests as normal, those above the cutoff for both tests as demented, and those in between as cognitively impaired. For the FAQ and BOMC tests combined, sensitivity in *detecting dementia* is 64.4% and specificity in *detecting normals* is 100%. Again, sensitivity is very poor for *detecting cognitive impairment* (24.3%). The predictive values (classification accuracy) for the FAQ, the BOMC, and the FAQ/BOMC combined tests are 73%, 64% and 60% respectively, for the sample of 609 subjects.

Figure 1 shows the results when ML methods are applied to the BOMC and FAQ tests' data to classify DSM-IV dementia status. The x-axis represents the number of patient examples in the *training sample* and the y-axis represents the ML algorithm's classification accuracy. Each data point in the figure is the mean of 30 runs. There is one curve of points for each of the four ML systems. All the ML systems except IB1 approach similar predictive values (classification accuracies) of 84.3%. The maximum standard deviation was 0.64% for any point.

The classification accuracies of each ML method used in conjunction with the FAQ and BOMC test results from the 609 subjects with a training sample size of 449 subjects are shown in **Table 3**. All ML methods tested show Compared to the FAQ and BOMC classification accuracy using published cutoff criteria, the *C4.5 results* show a 27.8% improvement in detecting dementia and a 36.5% improvement in detecting cognitively impaired subjects. The price paid is a 29.1% decrease in correctly classifying normals. This price is acceptable

for screening purposes. Comparing the cutoff criteria results with those obtained using *C4.5 Rules* shows a 28.2% improvement in detecting dementia and a 39.5% improvement in detecting cognitively impaired subjects. The price paid is a 34.2% decrease in correctly classifying normals. Comparison with IB1 and Naive Bayes methods give similar improvements in the ability to detect cognitive impairment and dementia.

The two algorithms, *C4.5* and *C4.5 Rules*, also generate easily understandable descriptions of how to classify a subject's dementia status. Some of the conditional statements generated by *C4.5 Rules* for classifying dementia status appear in **Table 4**. The percentages listed in brackets are the percent accuracy of the given rule for a testing sample. The rules should be tested in sequence. All of the conditions for a given rule must be true for the rule to apply. If no rules are true, then the default class, *demented*, is predicted.

Discussion

Of reported studies of diagnostic accuracy for dementia using the cutoff criteria for the BOMC, Watson et al.[18] reported an 82% sensitivity for detecting dementia and an 88% specificity for detecting normals (predictive value = 85%) among 76 skilled nursing facility residents. Pfeffer[19] reported a 93% dementia detection sensitivity and a 68% specificity of detecting normals among 158 SDAT subjects and 419 community controls (predictive value = 74.8%). The lower classification accuracies we obtained using the published cutoff criteria for the BOMC (74%) may result from a more heterogeneous sample: for non-normal subjects, 35.4% have Alzheimer's Disease, 21% have Vascular dementia, 23.4% have a mixed etiology, and 20.2% have other causes. Since the BOMC is heavily-weighted towards memory, as its name implies, a more heterogeneous set of dementia etiologies which do not equally impair memory may explain the lower classification accuracy we obtained with the published criteria. We have obtained permission to analyze the national Alzheimer's registry database, CERAD, to determine whether the improvements obtained using ML methods in this study with a heterogeneous sample can be obtained with less heterogeneous samples.

Validity studies using the BOMC have lumped the normal and cognitively impaired categories together. The AHCPR recommends using a total BOMC score greater than 10 to indicate dementia. Using ML methods in conjunction with the BOMC and FAQ tests allow the formation of explicit criteria that can distinguish normal from cognitively impaired from demented subjects. The AHCPR committee has emphasized the importance of identifying subjects with cognitive impairment since they are a very high risk group amenable to early evaluation, differential diagnosis and intervention.

Do these results generalize to the population of all persons at risk for dementia? To the extent that this sample represents such a population, they do generalize. Relative to other reports on the validity of the BOMC, the heterogeneity of etiologies of dementia and cognitive impairment in our sample is more representative of what exists in the population. The best validation sample, of course will be a prospectively followed, random community sample.

How do ML methods improve predictive value over that obtained using conventional

criteria? ML methods use a variety of methods for learning about data. They perform a heuristic search over the data looking for regularities. In this study, each ML algorithm uses a different method for finding attributes that are predictive of dementia status. *C4.5* and *C4.5 Rules* search to combine attributes from the BOMC and FAQ data, which may not be easily visible, but nevertheless aid in classifying dementia status. *Naive Bayes* looks for regularities by examining the prior probabilities of attributes using Bayes theorem. *IB1* looks for examples in the training sample that look similar to unclassified cases to predict dementia status. Generally, Machine Learning systems create a more complex description of the dementia status, resulting in better discriminating power over conventional criteria.

The results from each ML method differed from each other since each ML method is specialized to work on domains with particular characteristics. The searching heuristics and learning methods implemented in the ML systems greatly influence their classification performance. Since each ML system applies a different method for learning dementia state, their classification accuracies will differ.

Conclusion

This analysis demonstrates that in the absence of delirium, the FAQ and BOMC tests, used in conjunction with ML methods improve the correct classification of cognitively impaired and demented subjects by 28-40% over the use of the FAQ and BOMC tests using published cutoff criteria. More importantly, the largest improvements in detection are for the very high risk group, *cognitively impaired but not sufficient to meet criteria for dementia.* The 29%-34% decline in correctly classifying normals is acceptable for most screening purposes.

Since the FAQ and BOMC tests require minimal training to give and perform, they can be used in variety of health care and possibly home settings. ML methods used with these two tests provide simple, unambiguous, logical statements (*see* **Table 4**) to classify a patient's dementia status more precisely and accurately than do published criteria.

Future work will use ML systems in conjunction with other AHCPR-recommended tests involved in the assessment of dementia to implement AHCPR guidelines in a straightforward fashion for use by subjects as well as clinicians in detecting early cognitive impairment and dementia. Future research will also focus on improving the ML classification accuracy of 61%-64% on the high risk cognitive impairment group.

Acknowledgment

We wish to thank James M. Chen and Brian Craft for their development of the Alzheimer's Intelligent Interface (http://www.ad.uci.edu).

Table 1
Sample Characteristics by Dementia Status

	normal	Cog_Imp	Demented	Total
Males	18	57	175	250
Females	24	50	285	359
Over Age 65	25	70	415	510
Under Age 65	17	37	44	98
> 12 Yrs Education	34	79	213	326
<= 12 Yrs Education	8	28	244	280
Skilled Jobs	38	62	166	266
Unskilled Jobs	1	5	70	76
# of Subjects	42	107	460	609

Table 2

BOMC Prediction of Dementia	DSM-IV dementia status			
	normal	Cog_Imp	Demented	Total
normal	**42**	99	119	260
demented	0	8	**341**	349
Total	42	107	460	609

FAQ Prediction of Dementia	DSM-IV dementia status			
	normal	Cog_Imp	Demented	Total
normal	**42**	87	79	208
demented	0	20	**381**	401
Total	42	107	460	609

BOMC and FAQ Combined Prediction of Dementia	DSM-IV dementia status			
	normal	Cog_Imp	Demented	Total
normal	**42**	80	34	156
Cog_Imp	0	**26**	130	156
Demented	0	1	**296**	297
Total	42	107	460	609

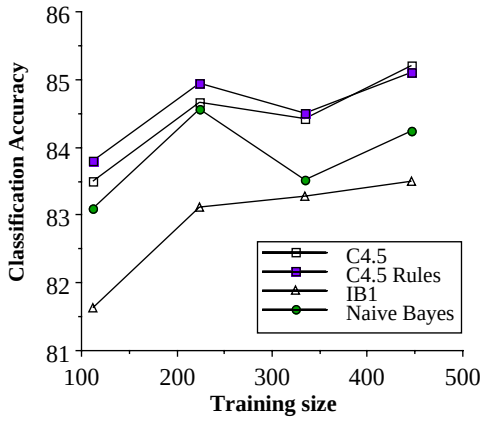

Figure 1. % Correct Classification of Dementia Status vs. Number of Learning Trials for Four ML Methods.

Table 3
Probability (%) of Correct Classification of Dementia Status by each ML method

	DSM-IV dementia status			
ML Method	Normal	Cognitively Impaired	Demented	Overall
C4.5	70.9	60.8	92.2	85.5
C4.5 Rules	65.8	63.8	92.6	85.9

Table 4. A sample of exemplary rules from *C4.5 Rules*.

```
Rule 1:

IF     count_20_backwards_to_1_score <= 1 AND  Recall_item1_score > 0 AND

     FAQtotal <= 0.1 AND age > 69

THEN  class normal  [83.3%]

Rule 2:

IF     Total_BOMC_Error_Score > 0.178571 AND   FAQ1 > 0 AND

     FAQtotal > 0.1 AND age > 55

THEN  class demented  [98.6%]

Rule 3:

IF     Total_BOMC_Error_Score <= 0 AND FAQtotal > 0.1 AND age <= 72

THEN  class cognitive_impairment  [85.2%]

...

...

Default class: demented
```

References

[1] G McKhann, D Drachman, M Folstein, R Katzman, D Price, and EM Stadlan. Clinical diagnosis of alzheimer's disease: report of the nincds-adrda work group under the auspices of department of health and human services task force on alzheimer's disease. *Neurology*, 34(7):939–44, Jul 1984.

[2] AM Kazee, TA Eskin, LW Lapham, KR Gabriel, KD McDaniel, and RW Hamill. Clinicopathologic correlates in alzheimer disease: assessment of clinical and pathologic diagnostic criteria. *Alzheimer Disease and Associated Disorders*, 7(3):152–64, Fall 1993.

[3] RS Hoffman. Diagnostic errors in the evaluation of behavioral disorders. *JAMA*, 248:225–8, 1982.

[4] DW O'Connor, A Fertig, MJ Grande, JB Hyde, JR Perry, MO Roland, JD Silverman, and SK Wraight. Dementia in general practice: the practical consequences of a more positive approach to diagnosis. *Br J Gen Pract*, 43:185–8, 1993.

[5] RL Ernst and JW Hay. The u.s. economic and social costs of alzheimer's disease revisited. *American Journal of Public Health*, 84(8):1261–4, Aug 1994.

[6] DP Lubeck and T Mazonson PD, Bowe. Potential effect of tacrine on expenditures for alzheimer's disease. *Medical Interface*, 7(10):130–8, Oct 1994.

[7] TF Williams and PT Costa. Recognition and initial assessment of alzheimer's disease and related dementias: Clinical practice guidelines. Technical report, Department of Health and Human Services, 1995.

[8] RI Pfeffer, TT Kurosaki, CH Harrah, JM Chance, and S Filos. Measurement of functional activities in older adults in the community. *J Gerontology*, 37:323–9, 1982.

[9] GG Fillenbaum, A Heyman, WE Wilkinson, and CS Haynes. Comparison of two screening tests in alzheimer's disease. the correlation and reliability of the mini-mental

state examination and the modified blessed test. *Archives of Neurology*, 44(9):924–7, Sep 1987.

[10] KA Welsh, N Butters, RC Mohs, D Beekly, S Edland, and G Fillenbaum. The consortium to establish a registry for alzheimer's disease (cerad. part v. a normative study of the neuropsychological battery. *Neurology*, 44(4):609–14, Apr 1994.

[11] American Psychiatric Association, Washington, D. C. *Diagnostic and Statistical Manual of Mental Disorders*, 4 edition, 1994.

[12] JR Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, Los Altos, California, 1993.

[13] R Duda and P Hart. *Pattern classification and scene analysis.* John Wiley & Sons, New York, 1973.

[14] D Aha, D Kibler, and M Albert. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[15] R , Kohavi, G John, et al. Mlc++: a machine learning library in c++. In *Tools for Artificial Intelligence Conference.* IEEE Computer Society Press, 1994.

[16] E Charniak and D McDermott. *An introduction to artificial intelligence.* Addison-Wesley Publishing Company, Menlo Park, California, 1985.

[17] R Gray. *Entropy and Information Theory.* Springer-Verlag, New York, 1990.

[18] YI Watson, CL Arfken, and SJ Birge. Clock completion: An objective screening test for dementia. *Journal of the American Geriatrics Society*, 41:1235–40, 1993.

[19] RI Pfeffer. A social function measure in the staging and study of dementia. In *Aging, Health, and Healing*, pages 459–74. Springer, New York, 1995.