Dementia Screening with Machine Learning Methods

William R. Shankle [1] Subramani Mani [2]
Michael J. Pazzani [3] Padhraic Smyth [4]

29 April 1997

[1]has joint appointments with the Departments of Neurology and Information and Computer Science, University of California at Irvine. His research interests include cognitive modeling and machine learning methods applied to developmental and degenerative conditions of the human brain.

[2]is a postgraduate researcher in the Department of Information and Computer Science, University of California at Irvine. He is interested in applications of machine learning, knowledge discovery & data mining, and bayesian networks to medicine.

[3]is an Associate Professor and Chairman of the Department of Information and Computer Science, University of California at Irvine. His research interests include inductive learning, theory revision and information retrieval using machine learning.

[4]is an Assistant Professor in the Department of Information and Computer Science, University of California at Irvine. His research interests are knowledge discovery and data mining, statistical pattern recognition and machine learning.

Abstract

Machine learning algorithms were applied to an electronic patient database generated by the UC Irvine Alzheimer's clinic to learn the simplest and most accurate patient parameters that would discriminate 244 very mildly demented from 198 normally aging subjects. Attributes included age, sex and education plus responses to the Functional Activities Questionnaire, the Mini-Mental Status and Blessed Orientation, Memory and Concentration tests. The Machine learning algorithms included decision tree learners(C4.5, CART), rule inducers(C4.5Rules, FOCL) and naive Bayes. Stepwise Logistic Regression was used to compare results. The sample was randomly split into training and testing sets, and the results were validated over 30 runs. Although the Functional Activities Questionnaire has been used since 1980, the Machine learning algorithms were the first to identify that a single attribute, measuring forgetfulness, equaled or exceeded the accuracy of any other scoring method of this test. Post hoc inspection of the odds ratios obtained by Stepwise Logistic Regression confirmed this finding. The application of Machine learning has identified an extremely simple, yet accurate screen for very mild dementia.

1 Introduction

With the advent of Electronic Medical Records (EMR) the size of patient databases has greatly increased. Traditional methods are not sufficient to analyze and interpret the enormous amounts of data being generated. Machine learning (**ML**) techniques can be valuable in this context for extracting useful information from large medical data sets. In this paper, we apply ML methods to the detection of the earliest stages of dementia due to Alzheimer's disease and other causes.

Machine learning can generate classification rules where the data include the known classification of each case. The application of ML methods in the domain of medicine has been relatively infrequent, partly because of past difficulties in accessing medical data electronically. Artificial intelligence approaches to medicine started with knowledge-based systems, constructed from knowledge provided by human experts, not data. Beginning with the expert systems of the seventies (MYCIN [Shortliffe, 1976], PUFF [Aikins et al., 1982]), followed by Bayesian systems of the late eighties and early nineties (ACORN [Wyatt, 1989], PATHFINDER [Heckerman et al., 1992]), these knowledge-based systems generated much enthusiasm. But there are very few such actual systems in routine clinical use. Another approach starting in the mid eighties sought to make use of real data and a domain model for knowledge acquisition and rule learning[Cestnik et al., 1987],[Michalski et al., 1986], [Lavrác and Mozetíc, 1992]. KARDIO [Brátko et al., 1989] is an expert system for evaluation of electrocardiograms based on this approach. With increasing availability of electronic medical records, machine learning has the potential to become a valuable adjunct to clinical decision-making. There has been some recent effort in this direction[Ohmann et al., 1995].

Dementia is defined as multiple cognitive impairments with loss of related functional skills without altered consciousness. A simple, unobtrusive method for detecting dementia early in the disease's course would help get patients to seek early evaluation and treatment, resulting most probably in preserving the existing quality of life and reducing the financial burden to family and health care providers. The Agency for Health Care Policy Research (**AHCPR**) clinical practice guidelines for the assessment and recognition of Alzheimer's disease and related disorders [Williams and Costa, 1995] recommends two simple tests, the Functional Activities Questionnaire (**FAQ** [Pfeffer et al., 1982]), and the six-item Blessed Orientation, Memory and Concentration test (**BOMC** [Fillenbaum et al., 1987]), to screen for dementia after excluding delirium and depression. We recently reported that the use of ML methods in conjunction with the FAQ and the BOMC markedly improved sensitivity in detecting dementia in a sample of 609 normal, cognitively impaired, and demented subjects when compared with published scoring criteria [Shankle et al., 1996].

In this paper, we focus on discriminating the effects of normal aging on cognition from the very early stages of dementia because early detection is po-

tentially very important for improving quality of life, and reducing total health care costs to family and society. To do this, we used the AHCPR-recommended screening instrument, the FAQ, plus the Folstein Mini-Mental Status Exam [Folstein et al., 1975] and two items from the BOMC not included in the MMSE (we henceforth use **MMSEBOMC** to denote this particular combination of tests) in conjunction with several ML methods and stepwise logistic regression, and compared these results to those using published scoring criteria for the same set of data from the same set of subjects. Other items of the BOMC did not need to be considered in addition to the MMSE since the rest of the BOMC is a subset of the MMSE. See Section 2 for a description of these tests.

2 Description of the data set

Sample Description

The total sample consisted of the initial visits of 198 cognitively normal and 244 cognitively impaired or very mildly demented (Clinical Dementia Rating Stage ≤ 0.5) subjects seen at the University of California, Irvine Alzheimer's Disease Research Center (ADRC). Patients received a complete diagnostic evaluation consisting of patient and caregiver interviews, general physical and neurological exam, two hours of cognitive testing including the CERAD [Welsh et al., 1994] neuropsychological battery and other selected tests, routine laboratory testing for memory loss, and magnetic resonance neuroimaging with or without single photon emission with computed tomography. Control subjects 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. The number of subjects available for the various analyses varied somewhat because of missing data. The sample sizes for each screening test appears in Table 1.

Classification of Dementia Status

The diagnosis of dementia status, using DSM-IV criteria [DSM-IV, 1994], 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, 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. They were also classified by dementia severity using standard criteria for the Clinical Dementia Rating Scale (**CDRS** [Morris, 1993]), in which 0 = normal,

Table 1: Characteristics of the UCI ADRC Sample of this study

Attrib	Normal			Impaired			Total		
	N	M	SD	N	M	SD	N**	M	SD
Age*	196	67.2	11.8	278	68.2	10.9	474	67.6	11.3
% Female*	198	71	–	274	43	–	472	59	–
Yrs Education	140	15.0	2.7	274	15.3	3.2	414	15.2	3.0
FAQ	137	0.2	0.8	211	7.6	6.2	348	5.1	6.1
MMSE	198	29.2	0.9	227	24.8	5.5	425	26.6	4.8

* T-test for normal vs. impaired groups (unpaired samples with unequal variances) was significant at $P < 0.001$

** Sample size varies due to missing data

0.5 = questionably or very mildly demented, and 1–5 indicate increasing severity of dementia. Control subjects showing cognitive impairment or very mild dementia (CDRS ≤ 0.5) were included in the cognitively impaired/very mildly demented sample, which we will refer to as the *impaired* group. Patients who tested normally were included in the cognitively normal sample; subjects with delirium were excluded from the analysis. Table 1 shows the sample characteristics.

FAQ and MMSEBOMC tests

The FAQ consists of ten questions about basic and more complex activities of daily life. The total score ranges from 0 (normal) to 30 (severely disabled). The answers to these questions were extracted from the UCI ADRC relational database of over 1,200 variables per subject-visit to compute the FAQ total and item scores. The AHCPR recommends using total FAQ scores of 9 or higher for detecting impairment. Pfeffer[Pfeffer et al., 1982] found a total FAQ score of 5 or higher to be most sensitive as a second stage screen in discriminating normal vs. questionably demented subjects. We examined the sensitivity and specificity of total FAQ scores from 1 to 30 without ML methods. With ML methods, we used age, sex, education, and all FAQ attributes with and without the FAQ total score. The description for how these runs were performed is in the Machine learning methods section.

The MMSEBOMC consists of 11 questions from the MMSE regarding orientation for time and place, registration, attention, short-term recall, language, and drawing, plus two questions from the BOMC test (recall of an address and number of trials to correctly repeat the address twice), which we were added because of their potential sensitivity in detecting early dementia. To avoid overfitting the data, we held the proportion of attributes relative to group sample

size to about 10% by aggregating the individual MMSE attributes reflecting short-term recall, orientation to time, and orientation to place into three aggregate attributes respectively. The MMSEBOMC attributes therefore consisted of the three MMSE aggregate attributes, individual MMSE attributes reflecting registration, attention and drawing, and the two BOMC attributes. These attributes plus age, sex and education were used with ML methods to classify normal and impaired subject samples. The MMSE ranges from 0 (severely impaired) to 30 (no impairment). The occurrence of dementia increases with advancing age and decreases with increasing educational level. Depending upon a subject's age and education, a total MMSE score of 24 or higher is used to classify a subject as normal [Oconnor et al., 1989, Crum et al., 1993]. We examined the sensitivity and specificity of total MMSE scores from 1 to 30 without ML methods.

3 Methods

Specific algorithms

We concentrated on decision tree learners, rule learners and the naive Bayesian classifier. Decision trees and rules generate clear descriptions of how the ML method arrives at a particular classification. The naive Bayesian classifier was included for comparison purposes. MLC++ (Machine Learning in C++) is a software package developed at Stanford University [Kohavi et al., 1994] which implements commonly used machine learning algorithms. It also provides standardized methods of running experiments using these algorithms. C4.5 is a decision tree generator and C4.5Rules produces if-then rules from the decision tree [Quinlan, 1993]. Naive Bayes is a classifier based on Bayes Rule. Even though it makes the assumption that the attributes are conditionally independent of each other given the class, it is a robust classifier and serves as a good comparison in terms of accuracy for evaluating other algorithms [Duda and Hart, 1973]. FOCL [Pazzani and Kibler, 1992] is a concept learner which can incorporate a user provided knowledge of two types. First, when provided with a guideline or protocol directly, FOCL has the capacity for revision if the guidelines produce better classification rules than that produced from exploration of the data. Second, FOCL can accept information on each nominal variable indicating which values of the variable increase the probability of belonging to a class (such as impaired) and information on each continuous variable on whether higher or lower values of the variable increases the probability of belonging to a class. We call this, "constrained FOCL", in the experimental results. FOCL can also learn from the data only, without an initial input of constraints or guidelines. We call this, "unconstrained FOCL", in the experimental results. CART [Breiman et al., 1984] is a classifier which uses a tree-growing algorithm that minimizes the standard error of the classification accuracy based on a partic-

ular tree-growing method applied to a series of training subsamples. We used Caruana and Buntine's implementation of CART[Buntine and Caruana, 1992] (the "IND" package), and ran CART 10 times on randomly selected 2/3 training sets and 1/3 testing sets. For each training set, CART built a classification tree where the size of the tree was chosen based on cross-validation accuracy on this training set. The test accuracy of the chosen tree was then evaluated on the unseen test set.

Treatment of missing data

We used each ML algorithm's particular approach for handling missing data. In C4.5 missing attributes are assigned to both branches of the decision node, and the average of the classification accuracy is used for these cases. In the naive Bayesian classifier, missing values are ignored in the estimation of probabilities. In FOCL, any test on a missing value is treated as false. Therefore, it attempts to learn a set of rules that tolerates missing values in some variables. CART uses surrogate tests for missing values.

Generation of Training and Testing Samples

The samples for the FAQ, and MMSEBOMC ML and stepwise logistic regression analyses mostly overlapped but the sizes differed due to different patterns of missing data. For the FAQ there were 348 instances—137 cognitively normal and 211 impaired; for the MMSEBOMC there were 425 instances—198 normal and 227 impaired. We cross-validated the analytical results in the following manner. The complete sample of each screening test was used to randomly assign subjects to either the training or testing set in a 2/3 to 1/3 ratio. This was done 30 times with the complete sample of subjects to generate 30 pairs of training and testing sets.

ML Analyses

We ran experiments in which data from the FAQ and MMSEBOMC tests were used separately by each learning algorithm. The ML algorithms were trained on the training set and the resulting decision tree then classified the unseen testing set. The classification accuracy of each ML algorithm is hence the mean of the accuracies obtained for the 30 runs of the testing set. An example of one decision tree rule-set appears in Figure 1.

Stepwise Logistic Regression Analyses

Data from the FAQ and the MMSEBOMC were separately regressed against dementia status in the following manner (demographic attributes were included in both regressions). We applied Stata's [Stata, 1993] stepwise logistic regression

Figure 1: A C4.5Rule Set

Rule 1: age > 56 *and* job $> 2 \Rightarrow$ class **impaired**

Rule 2: money > 0 *and* forget $> 0 \Rightarrow$ class **impaired**

Rule 3: gender $= 0$ *and* age > 56 *and* forget $> 0 \Rightarrow$ class **impaired**

Rule 4: age > 56 *and* age ≤ 64 *and* forget $> 0 \Rightarrow$ class **impaired**

Rule 5: age > 73 *and* forget $> 0 \Rightarrow$ class **impaired**

Rule 6: forget $\leq 0 \Rightarrow$ class **normal**

Rule 7: Default $\Rightarrow$ class **impaired**

package (swlogis) to each randomly generated training set to obtain models consisting of the attributes' coefficients (odds ratios). We then tested each model's classification accuracy [1] with the testing set corresponding to the given training set. For the FAQ testing set, we assigned *lstat's* cutoff parameter value to be 39% for the proportion of normal cases; for the MMSEBOMC, the *lstat* cutoff parameter lstat was set to 47%. These cutoff values properly reflected the sample's prior probabilities. The means and standard deviations of the classification accuracy, sensitivity and specificity were computed for the 30 samples.

Using the same training and testing set pairs previously described, we also performed logistic regression using the FAQ's forgetting question as the only independent variable; this allowed a more direct comparison to the ML results obtained by CART. The means and standard deviations of the classification accuracy, sensitivity and specificity were computed for the 30 samples.

Nonsense Rules

It is possible for ML methods to generate a rule which makes no domain sense i.e. a "nonsense rule". The rule sets generated by the various ML methods were inspected for their clinical sense by an ADRC staff neurologist. After identifying the nonsense rules, we used FOCL to incorporate domain-specific knowledge that would prevent (constrain) such rules from occurring. We then compared classification performance of the constrained vs. unconstrained runs using FOCL to see how performance was affected. An example of a decision tree with a nonsense component follows:

[1] We used the *lstat, all* option of swlogis for this

```
forget > 0 (having trouble):
|   age <= 52 :
|   |   edulevel > 16 : normal (4.0)
|   |   edulevel <= 16 :
|   |   |   SHOP <= 0 (no trouble shopping): impaired (5.6)
|   |   |   SHOP > 0 (having trouble shopping): normal (2.0)
```

In this example, eight persons (5.6+2.0) were forgetful, 52 years old or younger, and had 16 or fewer years of education. Among them, those who could shop were classified as impaired while those who required assistance to shop were classified as normal: this is a *nonsense rule*, which arises because of insufficient examples covering the circumstances specified by the nonsense rule. As becomes apparent later, the appearance of such nonsense rules should encourage one to look for logical errors in the data, gather more data, constrain the ML method with domain-specific knowledge, or to search for a reduced rule-set using pruning techniques.

4 Results

We examined the sensitivity (probability of correctly classifying an impaired subject) and specificity (probability of correctly classifying a cognitively normal subject) for each ML run of the testing samples. For each run, the same statistics were also generated for the cutoff values of the total scores of the FAQ and MMSE without the use of ML methods, and for the stepwise logistic regression. Figures 2 and 3 respectively show the receiver operating characteristic (ROC) curves for the FAQ and MMSE total scores without ML methods, as well as the performance of the best results using various ML algorithms. In the ROC plot, the X-axis is the false alarm rate (1 minus specificity) and the Y-axis is the detection rate (sensitivity). Table 2 shows the classification results of each ML method and of published criteria for total MMSE and FAQ scores. A number of strategies were used to select an optimal decision tree for clinical use. We ordered pruned decision tree rule-sets by their frequency of occurrence across the different ML methods and runs. We examined the cross-validation procedure of CART, which selects the best single decision tree for a specified number of runs; we repeated this procedure 10 times. Each time, CART selected the same best decision tree.

With regard to possible biases between the normal and impaired samples, only age and sex showed statistically significant differences. However, the age difference between normal and impaired subjects was less than one year, which is not a clinically significant difference. Therefore, only sex showed a clinically and statistically significant difference, with a preponderance of females in the normal group. This possible bias can be evaluated by examining whether gender had a significant role in the ML or stepwise logistic regression results. Our analyses show that gender only affected the classification accuracy of the MMSEBOMC

Table 2: Sensitivity and Specificity of each Screening test by algorithm and published scoring criteria

FAQ (Normal = 137, Impaired = 211)							
%	CART	C45	C45R	FOCL	NB	FAQ>8	FAQ>4
Ss	93	92	89	94	67	20	49
Sp	80	78	79	80	97	99	96
Ac	88	88	85	89	83	51	68

MMSEBOMC (Normal = 198, Impaired = 227)						
%	C45	C45R	FOCL	NB	MMSE>24	MMSE>27
Ss	77	70	79	66	30	62
Sp	80	86	70	87	100	81
Ac	79	77	75	75	63	71

Ss – Sensitivity, Sp – Specificity, Ac – Accuracy, C45R – C45Rules, and NB – naive Bayes

test. For the FAQ test, figure 2 shows that the FAQ with ML methods outperformed the best of the published cutoff criteria for the total FAQ score. It is interesting to note that the cutoff score of 9 or higher, recommended by the AHCPR, has a considerably poorer sensitivity for discriminating very mildly demented from normal subjects (20%) than that obtained for the ML methods, FOCL, C4.5, C4.5Rules, and CART (93%).

One should also note that the number of questions needed to achieve accurate classification with ML methods is markedly reduced. In the case of CART, only one question is required (*"Do you require assistance remembering appointments, holidays, family occasions, or taking medications?"*). For the MMSEBOMC test, figure 3 shows that, when used with ML methods, classification accuracy is always higher than that obtained using any published cutoff values of the total MMSE score. Using constrained vs. unconstrained analysis of the data with FOCL, there did not appear to be a significant improvement in classification accuracy, but no nonsense rules were generated when constraining FOCL with domain-specific knowledge. Given the various search strategies for finding the best decision tree or rule-set for clinical use, all approaches converged on one main conclusion: the response to a single question from the FAQ test gave classification accuracy as good as any other rule set and better than any published criteria. This question, *"Do you require assistance remembering appointments, family occasions, holidays or taking medications?"*, we call the **forgetting rule**. All runs for all ML algorithms studied included this rule in the decision tree/rule-set; no other attribute was included in every decision tree/rule-set. Using CART's cross-validation procedure, this single rule decision tree was selected as the best tree on 10 out of 10 runs.

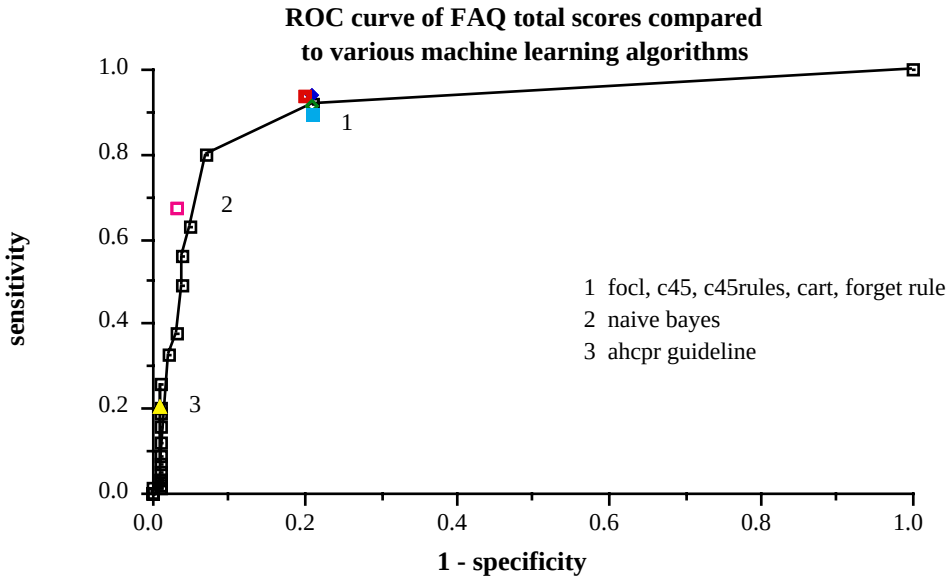

Figure 2: FAQ ROC

Logistic Regression Results

Stepwise logistic regression using the FAQ attributes gave a mean classification accuracy of $86.5 \pm 4.4\%$, which was similar to that obtained using ML methods (88%). The sensitivity ($84.1 \pm 6.8\%$) was lower than ML methods (93%), and the specificity ($88.8 \pm 5.3\%$) was higher than that obtained from ML methods (80%). The forgetting attribute had the largest odds ratio on 23 of 30 runs (11.9 ± 7.6), and was the only attribute included in all 30 models.

Logistic regression using the FAQ's forgetting attribute alone for the 30 randomly sampled training and testing sets gave a sensitivity of $92.7 \pm 2.0\%$ and specificity of $80.8 \pm 5.5\%$, which is significantly higher in sensitivity ($p < 0.00001$, two-sample t-test) and significantly lower in specificity ($p < 0.00001$, two-sample t-test) than that obtained by the stepwise logistic regression model.

For the MMSEBOMC, mean classification accuracy, $75.1 \pm 4.2\%$, sensitivity ($74.2 \pm 6.2\%$) and specificity ($76.5 \pm 6.8\%$) were not statistically different from those obtained by ML methods. In the logistic regression models, the attributes, *sex* and *repeating the months of the year in reverse order*, appeared in all 30 models; the attributes, *# of trials to obtain 2 correct repetitions of a previously unlearned address*, and *orientation to place*, appeared in 29 of 30 models; and the attribute, *delayed recall of a previously unlearned address*, appeared in 25 of

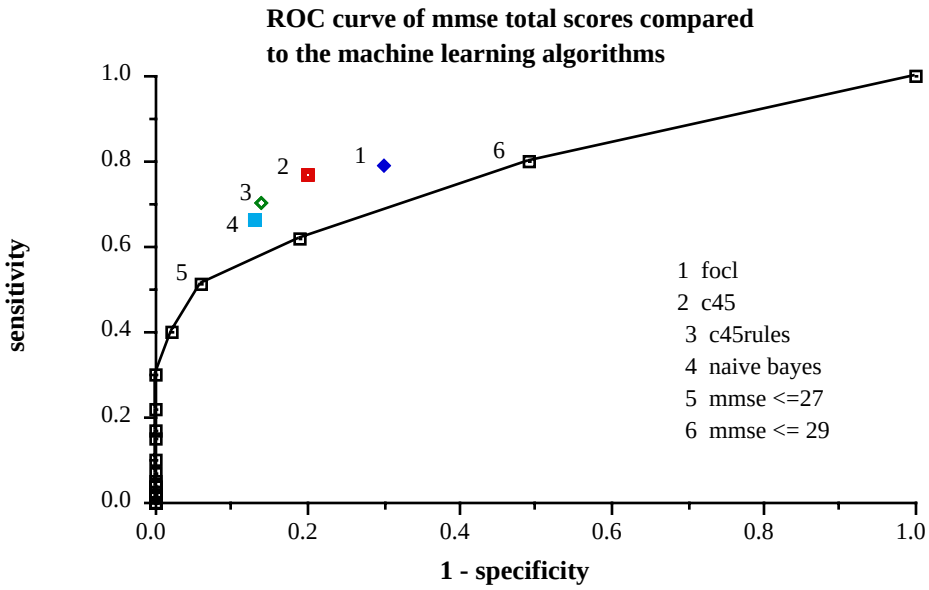

Figure 3: MMSE ROC

30 models. All other attributes appeared in less than 19 models. Among these five attributes, no one attribute had a distinctly larger odds ratio (range=1.4 to 6.4).

Comparison of sensitivity and specificity for the forgetting attribute logistic regression vs. the MMSEBOMC stepwise logistic regression using a two-sample t-test with unequal variances shows that the forgetting attribute alone gives statistically higher sensitivity ($p < 0.00001$) and specificity ($p < 0.0094$).

5 Discussion

There are four main findings of the present analysis. *First*, the ML methods can be interfaced with an electronic medical record system to learn directly from the data. The feasibility of this is also demonstrated by the work described in, for example, [Ohmann et al., 1995] and [Gierl and Stengel-Rutkowski, 1994]. This feature contrasts with that of knowledge-based systems, in which human experts design the decision rules and then test the data. Whereas humans usually select a few rules by which they make decisions, a machine can consider a larger number of rules if needed. When supplemented by a review of the ML-generated rules or by incorporation of domain-specific knowledge into the ML algorithm, specific rules that violate domain knowledge can be minimized, thus enhancing

the power and comprehensibility of rules obtained from ML methods. This approach also identifies subtle logical errors in the electronic medical record that could be overlooked. For example, after reviewing a nonsense rule using job performance as a criterion, we discovered that some normal subjects had misinterpreted the question about their ability to perform a job, answering that they could no longer perform their job because they had retired. In fact, they were fully able to perform their job given the need to do so. The inconsistency in the attribute values was discovered, and corrected. Re-running the ML algorithm verified that the nonsense rule had been eliminated by this correction of the data.

The *second* important finding of this paper is that ML methods used in conjunction with the MMSEBOMC test attributes outperform any published criteria for using total MMSE score to classify normal and cognitively impaired or very mildly demented subjects. They also do much better than any cutoff possible using the ROC curve. This supports the idea that some attributes of the MMSEBOMC are more important than others, and that the less important attributes may actually confuse classification. The findings of the logistic regression analyses also supported these conclusions.

The *third* important finding of this paper is clinical: when used with ML methods, a single question from the FAQ (the forgetting question) classifies cognitively normal aging subjects and subjects with very mild dementia as well as or better than any other combination of attributes from the FAQ and the MMSEBOMC with and without total score, and out-performs any of the recommended scoring criteria for the FAQ or the MMSE total scores. The logistic regression results confirm the ML finding that the FAQ forgetting attribute alone compared to the other best models from stepwise logistic regression of either the FAQ or the MMSEBOMC test gives the highest sensitivity (93%) for discriminating normal aging from the earliest stages of dementia. As expected, when both ML and logistic regression methods use only the information from the FAQ forgetting attribute to classify subjects, they give essentially identical sensitivity and specificity.

The similarity of the ML and logistic regression results also occurred in analyzing MMSEBOMC attributes. A recent study comparing Concept Formation ML methods with logistic regression also found highly similar classification results in predicting survival of injured patients entering the emergency department [Hadzíkadíc et al., 1996]. This suggests that other features of these methods besides classification accuracy are important in deciding which to use. In this study, ML methods more directly indicated the minimal set of attributes which accurately predicted normal aging vs. early stages of dementia. Inspection of the odds ratio of the FAQ forgetting attribute confirmed the ML finding. The presence of missing data also limited the number of cases available for analysis by logistic regression but not by ML methods. Unless one can easily perform calculations in a clinical setting, it is also easier to use the classification rules derived from ML methods than it is to work with regression coefficients

in classifying subjects. The FAQ forgetting attribute alone gave a higher specificity than the MMSEBOMC stepwise logistic regression results, but resulted in a lower specificity than that obtained from the stepwise logistic regression using the entire set of FAQ attributes. Therefore, compared to the best results obtained from all other permutations of the FAQ and MMSEBOMC test attributes, the forgetting attribute used alone for screening would incorrectly classify about 8% more normal aging individuals as cognitively impaired, and it would correctly classify about 6.6% more cognitively impaired persons as cognitively impaired. Given the ease and applicability of the FAQ forgetting attribute for screening, we think that the tradeoff for higher sensitivity is preferable, since one can apply a second screen in a clinical setting to eliminate normal aging individuals misclassified as impaired.

It is interesting to note that the AHCPR-recommended criteria for impairment using a total FAQ score of 9 or higher, is much higher than the score of a person answering positively only to the forgetting question (their FAQ total = 1–3 in that case). The higher total FAQ score recommended by the AHCPR is based on studies which included all levels of dementia severity. Using this criterion for the very mildly demented subjects in the present study resulted in only a 20% sensitivity, which implies that responses to other questions of the FAQ actually reduce the sensitivity for detecting very mild stages of dementia (compared to the forgetting rule alone). This is why inclusion of the total FAQ score as an attribute in the ML runs reduced the specificity and sensitivity when compared with the results obtained from analyses of the FAQ item attributes alone. The FAQ attributes therefore contribute unequally to dementia classification, with the forgetting question being the most contributory. This is our *fourth* significant finding.

Limits on Accuracy

Sample bias: The only demographic variable which differed to a clinically significant extent between normal and cognitively impaired subjects was sex. Since the decision rule sets obtained from the FAQ test plus demographic attributes rarely included gender in any of the ML runs, we conclude that FAQ decision rules are not biased by sample differences in gender. However, gender was a significant attribute in classifying subjects using the MMSEBOMC test. The findings here are restricted to the population represented, which consists of individuals, mostly over 65 years and with more than a high school education. Previous studies showing the insensitivity of the FAQ to educational level suggests that the results of this study also apply to persons 65 or over, regardless of education.

6 Conclusions

The dementia screening tests recommended by the Agency for Health Care Policy Research were analyzed with Machine Learning and Stepwise Logistic Regression methods. Compared to the most accurate cut-off criteria published for the total scores of these tests, Machine Learning methods increased the accuracy significantly, for the FAQ, by a wide margin of more than 30%. Furthermore, they reduced the number of test questions needed to obtain this accuracy to just one question. Stepwise Logistic Regression not only confirmed the Machine Learning results, but also assisted in the logical pruning of the decision trees through the inspection of the odds ratios of each attribute which participated in a rule-set. Despite the use of these tests in dementia evaluation for over 20 years, these findings have not been previously discovered, suggesting a useful role for Machine Learning in the evaluation of commonly used medical tests.

Also, Machine learning methods discovered subtle errors in the electronic medical record which were due to misinterpretation of what was being asked of the subject. The rule set derived from the full data can be used on paper or as software in various clinical settings to enhance the detection of very early stages of a dementing illness. This should result in less disability per patient and better quality of life for both caregiver and patient through early intervention. The utility of ML-derived protocols with some human supervision has general applicability to many important medical areas, including cancer, heart disease, and stroke.

acknowledgments

We thank professor Carl Cotman for helping establish a working relation with the AHCPR. This work was supported by the Alzheimer's Association Pilot Research Grant, PRG-95-161, *The Alzheimer's Intelligent Interface: Diagnosis, Education and Training.*

References

[Aikins et al., 1982] Aikins, J., Kunz, J., Shortliffe, E., and Fallat, R. (1982). Puff: an expert system for interpretation of pulmonary function data. Technical report, Stanford University.

[Brátko et al., 1989] Brátko, I., Mozetíc, I., and Lavrác, N. (1989). *KARDIO: A Study in Deep and Qualitative Knowledge for Expert Systems.* MIT Press, Cambridge, MA.

[Breiman et al., 1984] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth, Belmont.

[Buntine and Caruana, 1992] Buntine, W. and Caruana, R. (1992). *Introduction to IND Version 2.1 and Recursive Partitioning*. NASA.

[Cestnik et al., 1987] Cestnik, G., Konenenko, I., and Bratko, I. (1987). Assistant-86: A knowledge-elicitation tool for sophisticated users. In I.Brátko and N.Lavrác, editors, *Progress in Machine Learning*, pages 31–45. Sigma Press.

[Crum et al., 1993] Crum, R., Anthony, J., Bassett, S., and Folstein, M. (1993). Population-based norms for the mini-mental state examination by age and educational level. *JAMA*, 269(18):2386–2390.

[DSM-IV, 1994] DSM-IV (1994). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, Washington, D. C., 4 edition.

[Duda and Hart, 1973] Duda, R. and Hart, P. (1973). *Pattern Classification and Scene Analysis*. John Wiley, New York.

[Fillenbaum et al., 1987] Fillenbaum, G., Heyman, A., Wilkinson, W., and Haynes, C. (1987). 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.

[Folstein et al., 1975] Folstein, M., Folstein, S., and McHugh, P. (1975). Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–98.

[Gierl and Stengel-Rutkowski, 1994] Gierl, L. and Stengel-Rutkowski, S. (1994). Integrating consultation and semi-automatic knowledge acquisition in a prototype based architecture: Experiences with dysmorphic syndromes. *Artificial Intelligence in Medicine*, 6:29–49.

[Hadzíkadíc et al., 1996] Hadzíkadíc, M., Hakenewerth, B., Bohren, B., Norton, J., Mehta, B., and Andrews, C. (1996). Concept formation vs. logistic regression: Predicting death in trauma patients. *Artificial Intelligence in Medicine*, 8:493–504.

[Heckerman et al., 1992] Heckerman, D., Horvitz, E., and Nathwani, B. (1992). Towards normative expert systems: Part I The Pathfinder Project. *Methods of Information in Medicine*, (31):90–105.

[Kohavi et al., 1994] Kohavi, R., John, G., Long, R., Manley, D., and Pfleger, K. (1994). MLC++: A machine learning library in C++. In *Tools with Artificial Intelligence*, pages 740–743. IEEE Computer Society Press. Available by anonymous ftp from:
`starry.stanford.edu:pub/ronnyk/mlc/toolsmlc.ps`.

[Lavrác and Mozetíc, 1992] Lavrác, N. and Mozetíc, I. (1992). Second generation knowledge acquisition methods and their application to medicine. In Keravnou, E., editor, *Deep Models for Medical Knowledge Engineering*, pages 177–198. Elsevier, New York.

[Michalski et al., 1986] Michalski, R., Mozetíc, I., Hong, J., and Lavrác, N. (1986). The multi-purpose incremental learning system AQ15 and its testing application to three medical domains. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 1041–1045, Philadelphia, PA. Morgan Kaufmann.

[Morris, 1993] Morris, J. (1993). The clinical dementia rating (CDR): current version and scoring rules. *Neurology*, 43(11):2412–4.

[Oconnor et al., 1989] Oconnor, D., Pollitt, P., Treasure, F., Brook, C., and Reiss, B. (1989). The influence of education, social class and sex on mini-mental state scores. *Psychological Medicine*, 19:771–776.

[Ohmann et al., 1995] Ohmann, C., Yang, Q., Moustakis, V., Lang, K., and Elk, v. P. (1995). Machine learning techniques applied to the diagnosis of acute abdominal pain. In Barahona, P. and Stefanelli, M., editors, *Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine AIME95*, volume 934, pages 276–281. Springer.

[Pazzani and Kibler, 1992] Pazzani, M. and Kibler, D. (1992). The utility of knowledge in inductive learning. *Machine Learning*, (9):57–94.

[Pfeffer et al., 1982] Pfeffer, R., Kurosaki, T., Harrah, C., Chance, J., and Filos, S. (1982). Measurement of functional activities in older adults in the community. *Journal of Gerontology*, 37:323–9.

[Quinlan, 1993] Quinlan, J. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann, Los Altos, California.

[Shankle et al., 1996] Shankle, W., Datta, P., Dillencourt, M., and Pazzani, M. (1996). Improving dementia screening tests with machine learning methods. *Alzheimer's Research*, 2(3).

[Shortliffe, 1976] Shortliffe, E. (1976). *Computer-Based Medical Consultations: MYCIN.* Elsevier/North Holland, New York.

[Stata, 1993] Stata (1993). *STATA Release 3.1.* Stata Corporation, 6 edition.

[Welsh et al., 1994] Welsh, K., Butters, N., Mohs, R., Beekly, D., Edland, S., and Fillenbaum, G. (1994). 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.

[Williams and Costa, 1995] Williams, T. and Costa, P. (1995). Recognition and initial assessment of alzheimer's disease and related dementias: Clinical practice guidelines. Technical report, Department of Health and Human Services.

[Wyatt, 1989] Wyatt, J. (1989). Lessons learned from the field trials of ACORN, a chest pain advisor. In Barber, B., Cao, D., Qin, D., and Wagner, F., editors, *Proceedings MedInfo*, pages 111–115. Elsevier Scientific.