

Two-Stage Machine Learning Model for Guideline Development

Subramani Mani^a, William R. Shankle^{ab}, Malcolm B. Dick^c and Michael J. Pazzani^a
University of California, Irvine, California 62967, USA

^a Dept. of Information and Computer Science, ^b Dept. of Cognitive Science, ^c Dept. of Neurology

Abstract

We present a Two-Stage Machine Learning (ML) model as a data mining method to develop practice guidelines and apply it to the problem of dementia staging. Dementia staging in clinical settings is at present complex and highly subjective because of the ambiguities and the complicated nature of existing guidelines. Our model abstracts the two-stage process used by physicians to arrive at the global Clinical Dementia Rating Scale (CDRS) score. The model incorporates learning intermediate concepts (CDRS category scores) in the first stage that then become the feature space for the second stage (global CDRS score). The sample consisted of 678 patients evaluated in the Alzheimer's Disease Research Center at the University of California, Irvine. The demographic variables, functional and cognitive test results used by physicians for the task of dementia severity staging were used as input to the machine learning algorithms. Decision tree learners and rule inducers (C4.5, Cart, C4.5 rules) were selected for our study as they give expressive models, and Naïve Bayes was used as a baseline algorithm for comparison purposes. We first learned the six CDRS category scores (memory, orientation, judgement and problem solving, personal care, home and hobbies, and community affairs). These learned CDRS category scores were then used to learn the global CDRS scores. The Two-Stage ML model classified as well as or better than the published inter-rater agreements for both the category and global CDRS scoring by dementia experts. Furthermore, for the most critical distinction, normal vs. very mildly impaired, the Two-Stage ML model was 28.1% and 6.6% more accurate than published performances by domain experts. Our study of the CDRS examined one of the largest, most diverse samples in the literature, suggesting that our findings are robust. The Two-Stage ML model also identified a CDRS category, Judgment and Problem Solving, which has low classification accuracy similar to published reports. Since this CDRS category appears to be mainly responsible for misclassification of the global CDRS score when it occurs, further attribute and algorithm research on the Judgment and Problem Solving CDRS score could improve its accuracy as well as that of the global CDRS score.

Key words: Machine Learning, Clinical Dementia Rating Scale, Dementia Staging, Data mining

Author for correspondence

Subramani Mani, Rm. 444, Department of Information and Computer Science, University of California at Irvine, Irvine, CA 92697. Email: mani@ics.uci.edu Ph: 949-824-1316 Fax: 949-824-4056

1. Introduction

The staging of dementia severity using the Clinical Dementia Rating Scale (**CDRS**) [1,2] is done by clinicians using a two-stage process. After conducting a 30 minute structured interview with a reliable informant, the clinician uses CDRS criteria to rate the patient's level of impairment for each of six categories measuring memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. The clinician then uses another set of CDRS criteria [1,3] to assign a global level of impairment (global CDRS score) based on the six CDRS category scores. Various studies have reported up to an 80% inter-rater reliability in assigning a global CDRS score to a set of patients [4,5,6]. The complexity of this clinical task suggests a Two-Stage machine learning model (see *Methods* section).

With the availability of disease-retarding therapies, assessing the severity of a patient's dementia has become increasingly important to establish drug efficacy as well as to determine prognosis for family planning of patient care. Proper treatment can reduce the total cost of Alzheimer's disease by at least 17% [7,8], and early detection of dementia and correct staging of the severity of dementia is critical to optimizing quality of life and reducing costs. Although the CDRS is widely used in academic and research settings, it has not been useful in community practice because of the time and expense of a structured patient interview as well as the complexity of the logic for CDRS category and global scoring. Previous studies have identified flaws in the logic of the CDRS criteria used [9], and the Alzheimer's Disease Cooperative Study Unit (**ADCSU**) issued a revised CDRS [10] that corrected some, but not all, of these errors.

These shortcomings reflect the current methodology for guideline generation. Initially, paper-based guidelines are generated by a committee of experts. The evidence is assembled from population-based studies, case-control studies and meta-analyses. The algorithms for decision making are then rendered as flow charts, *if.. then* rules, or decision tables. The guideline implementers then convert these guidelines to software or other suitable format depending on the practice setting of the guideline user. Many of the logical flaws and shortcomings of the guidelines are detected at this late stage [11].

Knowledge Discovery from Databases (**KDD**) incorporates a sequence of steps such as preprocessing of data, data mining, model selection, post-processing and evaluation for finding useful structure from data. See [12] for a detailed discussion of these components. The Machine learning (**ML**) approach used here as the data mining step of KDD, offers an opportunity to learn the guidelines from the data collected in a hospital or community setting. By focussing on simpler models, some of the inadequacies resulting from current strategies for guideline generation could be addressed. Recently, ML techniques have been explored for disease screening [13], differential diagnosis [14,15], and other outcome measures [16,17].

If a method existed by which the CDRS could be used in community settings, given the time and monetary constraints of today's health care settings, the overall quality of dementia care could be improved. ML can facilitate clinical decision making by identifying the most relevant pieces of information among a larger set of data and provide them in the form of a decision tree or rule set for human use. We have previously used machine learning (ML) algorithms to assess dementia severity [18]. In the present study, we examine the accuracy and utility of ML algorithms in classifying CDRS category and global scores for 678 subjects ranging from normal aging to severe dementia. Given a maximal inter-rater reliability of approximately 80% [4,5,6], it may be possible to construct a set of easily used rules with similar or better accuracy. This will enable non-professional personnel to assign category and global CDRS scores with comparable accuracy, thereby making the use of the CDRS in community clinical settings practical.

2. Methods

2.1 Sample description

The sample consisted of the initial visits of 678 subjects seen at the University of California, Irvine Alzheimer's clinic between 1988 and 1996. Subjects ranged from normal aging to severely demented and were carefully assessed and diagnosed using a standard, four hour battery of tests which included the CERAD protocol for diagnosing Alzheimer's Disease [19]. Each subject has approximately 1,200 variables collected and entered into an electronic medical record for clinical research and patient care purposes.

2.2 Clinical derivation of the CDRS category and global scores: the clinicians’ computerized scoring algorithm

In order to eliminate a variability of at least 20% among raters assigning the CDRS global and category scores using the ADCSU criteria [1], we spent two years developing and validating a computerized scoring algorithm (CSA) of the CDRS category and global scores. Although this algorithm eliminates inter-rater variability and resolves ambiguities in the interpretation of the global and category CDRS scores not addressed by the ADCSU criteria, it is not practical for general use because of the time and cost required in gathering the necessary attributes. The algorithm is therefore a useful gold standard for developing less costly methods of dementia severity staging than clinical testing currently required for its use. We derived the computerized scoring algorithm for the CDRS category and global scores from the original variables collected on each of 352 subjects consecutively seen at the Alzheimer’s clinic. Because the ADCSU criteria for assigning the CDRS category and global scores are in some cases still ambiguous, we have resolved these ambiguities to be consistent with the consensus impression of the staff neurologist and neuropsychologist. About 100 subjects were needed to refine the CSA so that when discrepancies between clinician and computerized assignments of a CDRS category or global score occurred, the clinician changed their score in favor of that assigned by the CSA. We examined another 252 subjects and found that no further disagreements occurred between the CDRS global and category scores derived by staff experts vs. CSA that would have required us to further refine the algorithm. We then used this CSA to assign CDRS category and global score to all 678 subjects in the present study. These CDRS global and category scores derived by the CSA therefore have zero inter-rater variability, which eliminates one source of possible confusion in interpreting the reason for a difference in classification of a CDRS category or global score between human experts and ML-derived rules. Table 1 shows the distribution of these CDRS global and category scores derived by the computerized scoring algorithm for our sample.

2.3 Sample partitioning for Machine Learning experiments

The original subject data used for the present ML analysis were transferred from the center’s relational database using standard SQL scripts into formats acceptable to the ML

algorithms. The whole sample was randomly allocated into two equal partitions of 339 subjects (P_1 and P_2). The partition P_1 was used to learn the ML models to compute the CDRS category scores. Because CART generated simpler decision trees without a significant decrease in accuracy compared to the other ML methods tested, we selected the "best" CART model for each of the six CDRS categories (see section below for details). These models were used for computing the six CDRS category scores of each subject in the partition P_2 . These CDRS category scores were then used to learn the ML models for computing the CDRS global scores. For each partition P_1 and P_2 , we randomly assigned subjects to a training set or a testing set in a 2/3 to 1/3 ratio. This process was done 20 times to generate 20 pairs of training and testing sets for both P_1 and P_2 for mining by the ML algorithms. Figure 2 summarizes the experiments performed in classifying the global and category CDRS scores in order to evaluate the effects of Machine Learning (see the section, *Machine Learning classification of the CDRS global and category scores*, below for additional details).

2.4 Attributes used in the Machine Learning analysis

The attributes used in the ML analysis consisted of educational level, orientation and short term memory from the Mini-Mental Status Exam (MMSE) [20], the total scores from the CERAD Delayed Recall Test [21] and the Wechsler Memory Scale for Visual Reproductions [22], and total scores measuring judgment and reasoning from the WAIS-R Information and Similarities subscales [23]. Also included were the examiner's estimation of the patient's insight, and caregiver ratings concerning activities of daily living in the areas of job, hobbies, community activities, household activities, finances, and personal care. These were the same attributes that had been used to derive the gold standard global CDRS score using the clinicians' CSA.

2.5 ML model

The classical ML model is described over a feature (attribute) space X and a class (outcome variable) C . In this model we learn a function $f(X)$ which is predictive of C , i.e. $f(X) \Rightarrow C$. If there are irrelevant attributes, one could do attribute selection to obtain X'

such that X' is a proper subset of X . The attribute selection could be done by an expert, or by using feature selection algorithms. This feature subset X' could be input to ML algorithms to learn a new function, $f(X')$, improving or retaining the classification accuracy of the earlier function using all the attributes. The whole sample (instance space) is randomly partitioned into a training set and a test set. The training set is used to learn the model (function) and the test set is used for evaluation. This methodology increases the likelihood that the learned model will be valid on future instances representative of the sample space. Furthermore, this process of model learning and testing is repeated on different random splits of the sample to obtain more reliable estimates.

2.6 *The Two Stage model*

The Two Stage ML model is an extension of the classical ML model where a new set of attributes Z are learned from the existing set X . For this learning model to be feasible, two conditions should be met. Firstly, the domain being modeled should support a set of intermediate concepts represented by the set Z , where Z is not a subset of X , and there exists a mapping from X to Z such that each Z_i is dependent on a subset of X . Note that if Z is simply a subset of X , then the problem reduces to a feature selection task. Second, to facilitate learning of the set Z , gold standard labels should be available. This could be provided by domain experts. The underlying assumption here is that the function $f(Z) \Rightarrow C$ is a better model than $f(X) \Rightarrow C$. There are various advantages for such a model. The two-stage model is more expressive as it tracks intermediate concepts, and the decision pathway from the starting set of features to the final solution is made explicit. From a cognitive perspective, if domain experts follow such a decision making process, it is likely that such models will be preferred as they are more understandable. The availability of intermediate concepts in Two-Stage learning also facilitates easier model assessment and model improvement. If the accuracy of an intermediate concept $Z_i \in Z$, is not above a certain threshold, one could focus on revising that part of the feature space used to learn Z_i . As will be shown, the Z_i , “Judgment and Problem Solving” was one such intermediate concept that was poorly learned from its subset of X attributes.

2.7 Related work

The Two-Stage induction that we described in this paper is a form of constructive induction [24]. The original goal of constructive induction was to adding task specific features to the example representations to improve the accuracy and understandability of the learned concept. The induction of a hierarchy of intermediate concepts termed structured induction [25,26] is also similar. Here the domain expert is involved in attribute selection for decomposing a problem into sub-problems recursively and then induction is done bottom up. Our work differs from these previous ones in the following way. Earlier, while an expert provided rules for adding these features, in our work the expert provides classified training examples and the rules are learned. This new learned representation then serves as input to the next stage of induction. A similar idea was recently used in a mail filtering text classification program [27]. More recent work in constructive induction [28] removes the expert from the problem, and automatically constructs features that improve the accuracy of learned concepts. However, it is not clear that the experts in our domain would be willing to use such a system since the new concepts do not necessarily relate to named existing concepts used by the experts.

2.8 Specific ML algorithms

We concentrated on decision tree learners, rule learners and Naive Bayes. Decision trees and rules generate clear descriptions of how the ML method arrives at a particular classification. Naive Bayes was included for comparison purposes. MLC++ (Machine Learning in C++) is a software package developed at Stanford University [29] 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 [30]. 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 [31]. CART [32] is a classifier that uses a tree-growing algorithm which minimizes the standard error of the classification accuracy based on a particular tree-growing method applied to a series of training sub-samples. We used Caruana and Buntine's implementation of CART (the "IND" package) [33], and ran CART 20 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.

The intermediate concepts (CDRS category scores) and the outcome variable or class (global CDRS score) have a range of values between 0 and 3. Hence they could be considered *ordered*. However, as clinicians attach more importance and value to some distinctions (for example, normal aging versus very mildly demented), we have approached this as a classification problem from the ML perspective. But we provide a comparison with logistic regression models. We also note that as a future direction it might be worthwhile to evaluate regression tree models also for the task of dementia severity staging.

2.9 Machine Learning classification of the CDRS global and category scores.

Table 2 shows the mean classification accuracy of the six CDRS category scales, derived from the 20 training and testing sets analyzed by each ML algorithm. To see how different ways of applying Machine Learning to the task of classifying the category and global CDRS scores affect accuracy, we computed the global CDRS mean classification accuracy using Methods 2 through 5 described in Figure 2. Method 1 classifies both category and global CDRS scores using the CSA. For any given set of subject attribute data, Method 1 has no inter-observer variability, and is used as the gold standard for this study; classification accuracy is assumed to be 100%. The first row of Table 2 shows the mean classification accuracy for the CDRS global scale using method 4. The second row of Table 2 shows the mean classification accuracy for the CDRS global scale using method 2. Note that the ML models for the CDRS category scales were learned from the partition of the sample (P_1) that was not used for the ML analysis of the CDRS global score (P_2). Also, the last column of the second row of Table 2 shows the mean classification accuracy for the global CDRS score using method 3. This allows us to evaluate how well ML algorithms classify CDRS category scores independently of how well they classify CDRS global scores.

2.10 Model Selection

The three important considerations in model selection are its accuracy, comprehensibility and stability. There is no single formula applicable to all models. The selection criteria might vary depending on the domain and also the task the model is expected to perform. Comprehensibility is particularly important in medical domains. Let us analyze these components in some more detail.

2.10.1 Accuracy

Accuracy cannot always be represented just by a single measure. However, for some models, a total accuracy score might suffice. Generally at least two measures are used—sensitivity (predicted true positives divided by the total number of true positives) and specificity (predicted true negatives divided by the total number of true negatives). These two parameters are applicable for problems with a binary outcome. For instance, a model used for screening will lay more emphasis on sensitivity and might trade off specificity in the process. Accuracy above a threshold level chosen by the user, is a necessary prerequisite for a model to be acceptable. In the case of the CDRS, published inter-rater reliabilities are about 80% and ML models with accuracies of 80% or better do at least as well as the experts in dementia.

2.10.2 Comprehensibility

Comprehensibility of the model is of prime significance in many domains. This property of a model makes it adaptable and easily accessible for revision. There is consensus that simpler (smaller) models are more comprehensible, and researchers have recently focussed on additional factors such as adherence to domain constraints [34]. In this era of evidence-based medicine, with an attempt at making the practice of medicine much more uniform and objective, formulating effective protocols and guidelines that conform to domain knowledge is a high priority. For a guideline to be useful, comprehensibility is a necessary condition. There is preliminary evidence that models that do not violate the existing discipline rules are preferred by caregivers [34]. In our evaluation, the CART models in general scored high on comprehensibility, primarily because of their brevity.

2.10.3 Stability

Stability of a model can be looked at from a structural and functional point of view. Higher stability means less variance in the models generated from different partitions of the sample. When the majority of randomly generated training sets produce structurally similar models, it indicates that a stable model has been learned. Trees generated using C4.5 were more complex than those generated using CART. Because the decision nodes of the CART tree were also represented consistently in the top region of the C4.5 trees, these attributes, common to both treetops, are likely to make these models structurally stable. Otherwise, there would occur other highly correlated attributes that would confound the stability of the features residing in the root region of the CART and C4.5 trees. When highly correlated attributes do exist, feature selection could be used to filter them out.

The functional stability of structurally different models could be assessed by the variability in the classification of a representative test sample set apart for this task. Turney [35] discusses syntactic and semantic similarity measures (which correspond to structural and functional stability) for characterizing the stability of models. He has argued for semantic similarity measures to ascertain stability since they are less sensitive to superficial variations in representations. Likewise, they could prove useful in comparisons of stability across varied representations such as neural networks and decision trees. The reader is referred to [35] for a detailed discussion of model stability issues.

2.10.4 Adequacy

Adequacy of a model could be extrapolated as a function of its accuracy, comprehensibility and stability. It is mostly domain and user dependent. In the field of medicine, for a model to be effective and useful as a guideline, it has to score high on all these components. We could also consider *cost* of a model in human or dollar terms for ascertaining adequacy, in hospital or community practice settings. Adequacy is an area eschewed by ML research, which until formally addressed, will hold back its utility in real settings.

2.11 The “best” Machine Learning models for deriving CDR category scores.

There is no specific guideline to select the “best” rule set or decision tree from the ML models generated for the CDRS category scores. However, our previous work has shown that the conciseness and intelligibility of rules are major factors in their acceptance by healthcare professionals [34]. In particular, when there are clauses in a rule set which make no logical sense, the rule set is perceived as unacceptable by healthcare personnel. A second factor of primary importance in today’s general practice healthcare settings, which emphasize rapid processing of patients, is professional time. A rule set will not be used in general practice settings if physician time is required to collect the data. This means that information must be gathered by ancillary healthcare staff or by questionnaires filled out by patients or family members. If ancillary healthcare staff collect the data, time is still an important factor, while if patients or family provide the data without staff assistance, then time is much less important. Also, if healthcare personnel are involved in the data collection, then the cost is a product of the time required and the value of the participating staff’s time. It is clear that great weight should therefore be placed on informant-based data collection that minimizes healthcare staff time. Finally, mean classification accuracy must be acceptable to warrant the use of the rule set at all. In dementia severity staging, there are no clear cutoffs for acceptable accuracy, but in the case of the CDRS global score, domain experts generally show inter-rater reliabilities of about 80% [4,5,6]. This means that a classification accuracy of 75% or higher would be acceptable to domain experts if it saved them time and money. We therefore used a mean classification accuracy cutoff of 75% or higher to select the “best” models for assigning the CDRS global score. These three factors (rule set intelligibility and logic, professional staff time and cost per unit time, and classification accuracy) comprise, in our opinion, clinically practical criteria for selecting useful classification rule sets for the CDRS domain.

We decided to select one model to compute each of the six CDRS category scores to examine the feasibility of Machine Learning for generating guidelines. CART decision trees were used for this purpose because they were smaller than C4.5 trees while being similar in accuracy. One CART model was selected by the staff neurologist from among the 20 CART models generated for each CDRS category score based on the number of

attributes, their costs, the number of logically inconsistent rules in the decision tree, and overall accuracy for the CDRS category. An example of one of the CART decision trees selected is shown in Figure 1, for the CDRS category, Personal Care. The Personal Care sub-category is derived from five Activities of Daily Living (ADL) items and one Instrumental Activities of Daily Living (IADL) attribute. The ADL items were the ability to perform eating, dressing, grooming, bathing and elimination (bowel and bladder), and the IADL attribute included was pills (ability to take medication). The levels of performance assigned ranged from dependence to independence (Dependent = 3, Requires assistance = 2, Has difficulty but does by self = 1, and Normal = 0).

3. Results

3.1 CDRS global score classification by Machine Learning algorithms.

The performance of the ML algorithms in classifying the category and global CDRS scores is summarized in Table 2. The Two-Stage Machine Learning experiment (Method 2 of Figure 2) classified the global CDRS score (CDRS² of Table 2) 4.4% to 12.6% less accurately than when the CSA computed the CDRS category scores and the global CDR was learned (Method 4, Figure 2 and CDRS⁴ of Table 2). To avoid overfitting the data when both category and global CDRS scores were learned, and to obtain conservative and valid estimates, their sample partitions did not overlap. Both methods of learning the global CDRS scores did about as well as the inter-rater reliabilities published for experts in the literature [4,5,6]. Also, in our previous study [18], in which we used the same ML algorithms as in the present study to learn global CDRS scores directly from the original subject data (Method 5, Figure 2 and CDRS⁵ of Table 2), classification accuracy declined even further in comparison to the Two-Stage ML method. The Two-Stage ML approach is therefore superior to directly learning the global CDR score from the original attribute data.

The degree of misclassification of the global CDRS was examined with the mean confusion matrices obtained by the Naïve Bayes algorithm using, as input attributes, either the CART-learned or clinician-computed CDRS category scores (see Table 4). Overall, the global CDRS scores learned from the CART-learned CDRS category scores were frequently misclassified. However, misclassifications of the global CDRS score were

almost always only one level of impairment off. For the global CDRS scores learned by Naïve Bayes from the CART-learned CDRS category scores, there were proportionally fewer misclassifications of normal aging subjects (global CDRS = 0) than any other category. Since it is important that impaired subjects are not classified as normal, Naïve Bayes performed fairly well in that it misclassified only 8.8% of the *very mildly* demented subjects as normal, while misclassifying 20.6% of them as *mildly* demented. The Naïve Bayes algorithm using learned CDRS category scores misclassified the global CDRS score by one level of impairment for 30.4%, 20% and 0% of the mildly, moderately and severely demented subjects. Given the professional time and money saved using this Two-Stage ML approach, these levels of misclassification are quite acceptable. In community settings, their use would be a major improvement since most community practices use no method of staging dementia severity. In dementia research centers, their use would lead quickly to further improvements in classification with ML because of the larger sample size available to ML researchers.

3.2 CDRS category score classification by Machine Learning algorithms.

All CDRS categories except Judgment and Problem Solving obtained a classification accuracy above 75%. This is not surprising because Judgment and Problem Solving is the most difficult CDRS category to quantify. The ML algorithm with the best overall performance was Naïve Bayes. All the ML algorithms classified the CDRS categories, Personal Care, Orientation, and Memory with above 75% accuracy, while only C4.5 failed to reach this criterion for the CDRS category, Community Affairs. The most striking difference among the ML algorithms occurred in classifying the CDRS category, Home and Hobbies. Naïve Bayes was the only ML algorithm to obtain classification accuracy greater than 75% for Home and Hobbies, and was at least 10% more accurate than any other ML algorithm used in this study. This could be because it integrates all features rather than a subset used by decision tree learners, and was advantageous for learning the Home and Hobbies category score.

3.3 Logistic regression

Using 20 data sets consisting of the clinician-computed CDRS category and global scores for 452 randomly selected subjects (a 2/3 sample), we used a stepwise, ordinal logistic

regression to determine how much of the variability in the global CDRS scores could be accounted for. This procedure is more appropriate than logistic regression when the outcome variable (global CDRS) has more than two classes. In this case, there were five classes ranging from normal aging to severe dementia. The mean proportion of the variance explained by stepwise ordinal logistic regression was 63.7% (std. dev. = 1.75%). Interestingly, all CDRS category attributes statistically significant predictors of the global CDRS in all runs, and the CDRS category with the poorest explanatory power in all runs was judgment and problem solving.

Using the machine learned category CDRS scores and the clinician-derived global CDRS scores, we then used stepwise ordinal logistic regression to determine how much of the variability in global CDRS scores can be explained by the machine-learned CDRS category scores. We randomly sampled 2/3 of the 339 subjects available for this partition of the data set 20 times to create 20 data sets and then performed the regression. Because of missing data in the machine learned CDRS category scores the 2/3 samples ranged from 109 to 122 subjects. The average proportion of the variance of the global CDRS scores explained by this regression was 52.1% (std. dev. = 3.1%). In summary, regression methods did not perform as well in classifying these data as did machine learning methods.

4. Discussion

4.1 Salient Contributions

The most important finding of this study relevant to Machine Learning is the improved classification accuracy achieved through the use of an intermediate classification stage (learning the CDRS category scores). In Table 3, comparing the effect of learning the global CDRS score using an intermediate learning stage (learning the category CDRS scores) to that obtained without an intermediate learning stage shows a 1.9% to 11.3% improvement by using an intermediate learning stage.

The ML stage that learns the global CDRS score from the category CDRS scores classifies with similar accuracy (except in the case of CART) to that achieved by the perfected clinicians’ CSA when given the same learned category CDRS scores as input (See Table 3, last row).

The majority of the decline in global CDRS classification accuracy occurs at the stage of Machine Learning the CDRS category scores from the original subject data. In Table 3, a 21.6% reduction in classification accuracy occurs when the clinicians' CSA is given machine learned CDRS category scores. Reduction in classification accuracy also occurs when Method 2 (the Two-Stage learning experiment) is compared to Method 4 (Machine learning the global CDRS score using the category CDRS scores derived by the clinicians' CSA). Comparing these two methods shows that there is a 6.2% to 12.6% reduction in global CDRS classification accuracy due to the CDRS category scores during the Two-Stage learning experiment. Hence, improving the classification accuracy of the CDRS category score stage would boost global CDRS score classification accuracy.

Classification accuracies above 75% were obtained for learning five out of the six CDR categories (see Table 2). Judgment and Problem Solving was the only CDRS category that none of the ML algorithms classified very well. This actually parallels the inter-rater agreement of human experts [4]. The attributes used for computing the category of Judgment and Problem Solving had a disproportionately high rate of missing data that resulted in a missing code for 25% of instances for this particular category. This in turn affected the learning of global CDRS and hence further research to improve the accuracy of this category score is called for. Also, for the CDRS category, Home and Hobbies, only Naïve Bayes classified with accuracy above 75%. In the Two-Stage ML model, learning the global CDRS score from the category CDRS scores is extremely accurate (see Table 3, last row). Hence, the greatest improvement in classification accuracy will most likely result from improved ML models classifying the CDRS categories, Judgment and Problem Solving, and Home and Hobbies. This data mining research has therefore identified the specific areas where further improvements in staging dementia severity can be obtained automatically.

How acceptable is the classification accuracy obtained through Machine Learning? Table 2 shows that learning the global CDRS score directly from the original subject data results in classification accuracy between 64% and 72%, which are probably not clinically acceptable, compared to the published inter-rater reliabilities of about 80% [4,5,6]. Introduction of an intermediate learning stage gives global CDRS classification accuracy

between 70.3% and 77.9%, which is reasonably close to performances achieved by domain experts. Even if classification accuracy is not improved by Machine Learning, there are still advantages to using the Two-Stage Machine Learning classification of the global CDRS score. These advantages are 1) professional time is not required to derive the global and category CDRS scores once the attribute data are collected; 2) the Machine Learning process identified where to focus further research to increase the ML classification accuracy of the global CDRS score; 3) the number of attributes required to classify the global and category CDRS scores is a substantially reduced subset of that required by domain experts, and eliminates certain costly tests for this purpose, such as the Wechsler Memory for Visual Reproductions Scale. Other advantages to using ML for classifying the global and category CDRS scores will depend upon the method of implementation. For example, using an electronic medical record, ML algorithms can be triggered once the requisite attributes have been entered. Also, ML algorithms could compute the CDRS scores while the attribute data are being collected to allow economy in attribute data collection, particularly for attributes, which are costly to collect (i.e. neuropsychological tests).

4.2 Inter-rater agreement (reliability) in dementia severity assessment

Inter-rater agreement refers to the match in global or category CDRS score assignment by raters (dementia experts or other specifically trained healthcare personnel). In their study of 25 patients with five reviewers, Burke et. al. report an inter-rater agreement of 80% for the global CDRS score [4]. Another study which focussed on nurse-physician and nurse-nurse inter-rater agreement using a sample size of 25 patients (20 having dementia resulting from AD and 5 healthy controls) obtained 80% to 81% for the global CDRS score, and 73% to 81% for CDRS category scores [5]. In their recent study on inter-rater variability using 82 investigators, randomly divided into two groups, and asked to rate three patients who fulfilled clinical diagnostic criteria for AD, the overall inter-rater agreement was 83% [6]. One would think that their agreement figure of 83% would be an upper bound considering their small sample size of patients plus the fact that they all met the clinical profile of AD. In contrast, our sample size was considerably larger ($n = 678$) and more complex in composition (included dementia of various origins—Alzheimer's disease, vascular dementia, depressive pseudo-dementia, Lewy Body dementia, and mixed

dementia). Interestingly, published inter-rater agreements for normal (CDR = 0), and cognitively impaired (CDR = 0.5) subjects were only 66% [6], compared to 94.1% for CDR = 0, and 70.6% for CDR = 0.5, using the Two-Stage ML model.

The acceptability of the Two-Stage ML classification of the global CDRS can also be evaluated by examining Table 4. It is evident that the Two-Stage ML model never misclassifies global CDRS by more than one level of impairment. Furthermore, the Two-Stage ML model never classifies normal aging subjects as very mildly demented, and misclassifies very mildly demented subjects as normal aging only 8.8% of the time, which is quite good. Therefore, when the Two-Stage ML model misclassifies a subject's global CDRS score, a gross error in classification does not occur making it clinically useful.

How reliable are the CDRS category scores learned by the ML algorithms, when compared to expert raters? Table 2 (rows 5-10) gives the accuracy for the different CDRS categories—Memory, Orientation, Judgment and Problem Solving, Community Affairs, Home and Hobbies, and Personal Care. All of the ML algorithms gave an accuracy of between 60% and 70% for the category Judgment and Problem solving, whereas for the other categories it was much higher. This ML result compares favorably with published performances by dementia experts, showing an inter-rater agreement of 68% for the category of Judgment and Problem solving, 80 to 88% for the other categories [4], and 73% to 81% for all the CDR category scores [5]. For the CDRS category Judgment and Problem Solving, future directions to improve the accuracy will include feature selection to reduce the attribute space by removing any irrelevant attributes, plus examination of item responses instead of just the total scores. Another approach worth pursuing is to capture the interactions among the attributes using a Bayesian network architecture [36].

4.3 Has the Two-Stage Machine Learning method identified a clinically usable model to stage dementia severity?

The answer is a qualified yes. We have compared the value of the ML methods reported here to the published data on dementia experts scoring the global and category CDRS scores for predominantly AD subjects. Most of these studies are not amenable to Machine Learning because of their small sample sizes of about 25 subjects. The performance of the

Two-Stage ML methods we examined is based on a total sample of 678 subjects, making it the largest study of the CDRS in the dementia literature. For both category and global CDRS scores, the Two-Stage ML methods have identified models that classify at least as well as dementia experts in terms of their inter-rater reliability and degree of misclassification. In normal and very mildly demented subjects, which will increasingly become the most important categories to accurately classify (due to the benefits of early detection, diagnosis and treatment) they outperform dementia experts by 28.1% and 4.6% respectively. Furthermore, the Two-Stage ML models we examined reduce the information space needed to assess dementia severity. Even though the CDRS category, Judgment and Problem Solving, clearly needs to be improved in its classification accuracy, the Two-Stage ML methods used here can acceptably eliminate the need for domain experts to perform the task of applying the CDRS criteria to assess global and category-specific dementia severity. In short, the Two-Stage ML models developed here outperform any domain experts in the published literature.

Examining Table 2 (Method 2 of Figure 2) indicates which Two-Stage ML algorithms are most accurate for classifying the global and category CDRS scores. For the global CDRS score (see row 3 of Table 2) C4.5, C4.5Rules and Naïve Bayes classify best. For the Memory and Community Affairs CDRS scores, C4.5Rules, CART, and Naïve Bayes classify best. For the Orientation and Personal Care CDRS scores, all ML algorithms classify well. For the Judgment and Problem Solving CDRS score, C4.5 and CART classify best. For the Home and Hobbies CDRS score, only Naïve Bayes classifies well.

In terms of general utility in the community, several issues need to be addressed before a Two-Stage ML classification of the global and category CDRS scores can be used. The most important issue involves finding an alternate way of collecting information for attributes that are expensive to collect (i.e., judgment and reasoning, and delayed recall). At present, our neuropsychologist collects these attributes by administering the WAIS-R Information and Similarities subtests, and the CERAD Delayed Recall Test. If reliable measures of short term memory, and judgment and problem solving can be found that an informant, secretary, nurse-assistant or other less highly trained healthcare personnel can

administer, then attributes can be economically collected in community settings. That this may be possible is suggested by recent attempts to convert the CDRS instrument into a questionnaire that can be answered by a reliable informant without assistance [37]. The second issue that needs to be addressed is the method of implementing a Two-Stage ML model in community settings. Some offices will have computers, making software implementation feasible, while others will require a paper-based implementation. One should note that these issues are not barriers to implementing a Two-Stage ML model in dementia research centers; implementing ML models there would accelerate ML research resulting in development of better ML models for community settings provided the data can be pooled.

5. Conclusion

We have shown that Machine Learning using a Two-Stage learning model for dementia severity assessment significantly improves classification accuracy over a single stage ML model and outperforms stepwise ordinal logistic regression by a margin greater than 20%. We advocate this approach when the domain semantics permit learning of an intermediate stage. Once the intermediate concepts are learned, these become the feature (attribute) space for the second stage of learning. In the staging of dementia severity using the CDRS, clinicians use such a two-stage approach. Our study provides an automated expert-quality guideline generation schema for dementia staging, which simulates the Two-Stage problem solving methodology employed by experts. The process of learning these two stages had the additional advantage of identifying which parts (CDRS categories) of the intermediate stage require further improvement. The identification of a category (Judgment and Problem Solving) with low classification accuracy, by ML algorithms, emphasizes the importance of examining more data relevant to this category for improving overall global CDRS classification accuracy. This study illustrates the value of the Two-Stage ML model in the development and refinement of diagnostic guidelines.

Acknowledgements

We thank Cathy Blake and Stephen Bay for reviewing an earlier draft of this manuscript.

We are grateful to the detailed comments of the two anonymous reviewers and particularly thankful to the guest editors for their suggestions for revision.

References

- [1] J.C. Morris, L.A. Coben, E.H. Rubin et. al. Clinical Dementia Rating. In: Bergener M, Finkel SI, eds. *Treating Alzheimer's and other dementias: Clinical applications of recent research advances*. New York: Springer, 1995:338-346.
- [2] C.P. Hughes, L. Berg, W.L. Danziger, L.A. Coben and R.L. Martin. A new clinical scale for the staging of dementia. *British Journal of Psychiatry*, 140:566–72, Jun 1982.
- [3] J.I. Chui, D. Victoroff, W.J. Margolin, R. Shankle and R. Katzman. Criteria for the diagnosis of ischemic vascular dementia proposed by the state of California Alzheimer's disease diagnostic and treatment centers. *Neurology*, 42(3): 473–80, Mar 1992.
- [4] W.J. Burke, J.P. Miller, E.H. Rubin, J.C. Morris, L.A. Coben, J. Duchek, I.G. Wittels and L. Berg. Reliability of the Washington University Clinical Dementia Rating. *Archives of Neurology*, 45(1): 31–2, 1988.
- [5] M.M. McCulla, M. Coats, N. Van Fleet, J. Duchek, E. Grant, and J.C. Morris. Reliability of clinical nurse specialists in the staging of dementia. *Archives of Neurology*, 46(11): 1210–1, Nov 1989.
- [6] J.C. Morris, C. Ernesto, K. Schafer, M. Coats, S. Leon, M. Sano, L.J. Thal, P. Woodbury and the Alzheimer's Disease Cooperative Study. Clinical Dementia Rating training and reliability in multi-center studies: The Alzheimer's Disease Cooperative Study Experience. *Neurology*, 48:1508-1510, 1997.
- [7] M.S. Mittelman, S.H. Ferris, E. Shulman, G. Steinberg and B. Levin. A family intervention to delay nursing home placement of patients with Alzheimer disease. A randomized controlled trial [see comments]. *JAMA*, 1996 Dec 4, 276(21): 1725-31.
- [8] D.P. Lubeck, P.D. Mazonson and T. Bowe. Potential effect of tacrine on expenditures for Alzheimer's disease. *Medical Interface*, 1994 Oct 7(10): 130-8.
- [9] D.J. Gelb and R.T. St. Laurent. Alternative calculation of the global clinical dementia rating. *Alzheimer Disease and Associated Disorders*, 7(4): 202–11, 1993.
- [10] Alzheimer's Disease Cooperative Study Unit (Manual). *Assignment of Clinical Dementia Rating*. Jan. 1994.
- [11] R.N. Shiffman. Representation of clinical practice guidelines in conventional and augmented decision tables. *JAMIA*, 4:382-393, 1997.
- [12] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy. From data mining to knowledge discovery: An overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy eds. *Advances in Knowledge Discovery and Data Mining*, pp. 1-36, AAAI Press, Menlo Park, California, 1996.
- [13] W.R. Shankle, S. Mani, M.J. Pazzani, and P. Smyth. Detecting very early stages of dementia from normal aging with machine learning methods. In Keravnou, E., Garbay, C., Baud, R., and Wyatt, J., editors, *Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine, AIME97*, volume 1211, pp. 73–85. Springer, 1997.
- [14] S. Mani, W.R. Shankle, M.J. Pazzani, P. Smyth and M.B. Dick. Differential Diagnosis of Dementia: A Knowledge Discovery and Data Mining (KDD) Approach. *JAMIA supplement p875*, 1997. Full paper in extended proceedings (CD ROM).
- [15] C. Ohmann, Q. Yang, V. Moustakis, K. Lang, and P.J. van Elk. Machine learning techniques applied to the diagnosis of acute abdominal pain. In Pedro Barahona and

- Mario Stefanelli, editors, *Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine AIME95*, volume 934, pp. 276-281. Springer, 1995.
- [16] G.F. Cooper, C.F. Aliferis, R. Ambrosino, J. Aronis, B.G. Buchanan, R. Caruana, M.J. Fine, C. Glymour, G. Gordon, B.H. Hanusa, J.E. Janosky, C. Meek, T. Mitchell, T. Richardson and P. Spirtes. An evaluation of machine learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine* 9:107-138, 1997.
- [17] K.C. Abston, T.A. Pryor, P.J. Haug and J.L. Anderson. Inducing practice guidelines from a hospital database. *JAMIA supplement* 1997, 168-172.
- [18] W.R. Shankle, S. Mani, M.B. Dick and M.J. Pazzani. Simple Models for Estimating Dementia Severity Using Machine Learning. *Proceedings, MedInfo'98: 9th World Congress on Medical Informatics, Seoul, S. Korea, 1998.* (In Press)
- [19] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E.M. Stadlan. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA work group under the auspices of the department of health and human services task force on Alzheimer's disease. *Neurology*, 34(7): 939-44, Jul 1984.
- [20] M.F. Folstein, S.E. Folstein, and P.R. McHugh. Mini-mental state—A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3): 189-98, Nov 1975.
- [21] K.A. Welsh, N. Butters, R.C. 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.
- [22] G.P. Prigatano. Wechsler memory scale: a selective review of the literature. *Clinical Psychology*, 1978. Series title: *Archives of the behavioral sciences monograph series*; no. 54.
- [23] D. Wechsler. *Manual for the Wechsler adult intelligence scale.* Psychological Corp, New York, 1955.
- [24] R. Michalski. *Inferential Theory of Learning: Developing foundations for Multistrategy Learning.* In R. Michalski & G. Tecuci (eds) *Machine Learning, A Multistrategy Approach*, pp. 3-61, 1993.
- [25] A.D. Shapiro. *Structured Induction in Expert Systems.* Turing Institute Press and Addison Wesley Publishers, Menlo Park, California, 1987.
- [26] D. Michie. Problem decomposition and the learning of skills. In *Machine Learning: ECML-95, Lecture Notes in Artificial Intelligence*, 912, (eds. N. Lavrac and S. Wrobel), Berlin, Heidelberg, New York: Springer Verlag, pp. 17-31, 1995.
- [27] G. Boone. Concept features in Re:Agent, an intelligent E-mail Agent. *Proceedings of the Second International Conference on Autonomous Agents*, pp. 141-148, ACM Press, Minneapolis, 1998.
- [28] B. Zupan, M. Bohanec, J. Demsar and I. Bratko. *Machine Learning by Function Decomposition.* *Proceedings of the 14th International Conference on Machine Learning.* (pp. 421-429). Morgan Kaufmann, San Francisco, 1997.
- [29] R. Kohavi, George John, Richard Long, David Manley, and Karl Pflieger. *MLC++: A machine learning library in C++.* In *Tools with Artificial Intelligence*, pages 740-743. IEEE Computer Society Press, 1994.
- [30] J.R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, Los Altos, California, 1993.

- [31] R.O. Duda and P.E. Hart. Pattern Classification and Scene Analysis. John Wiley, New York, 1973.
- [32] L. Brieman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and Regression Trees. Wadsworth, Belmont, 1984.
- [33] W. Buntine and R. Caruana. Introduction to IND (Version 2.1) and Recursive Partitioning. NASA, 1992.
- [34] M.J. Pazzani, S. Mani and W.R. Shankle. Beyond concise and colorful: Learning intelligible rules. In The Third International Conference on Knowledge Discovery and Data Mining, pages 235-238. AAAI Press, Menlo Park, California, 1997.
- [35] P. Turney. Bias and the quantification of stability. Machine Learning, 20:23-33, 1995.
- [36] D. Heckerman. Bayesian networks for data mining. Data Mining and Knowledge Discovery, 1:79-119, 1997.
- [37] Clark, CM. and Ewbank, DC. (1996). Performance of the dementia severity rating scale: a caregiver questionnaire for rating severity in Alzheimer disease. Alzheimer Disease and Associated Disorders, 1996 Spring, 10(1): 31-9.

Figure 1. CART Tree for Personal Care

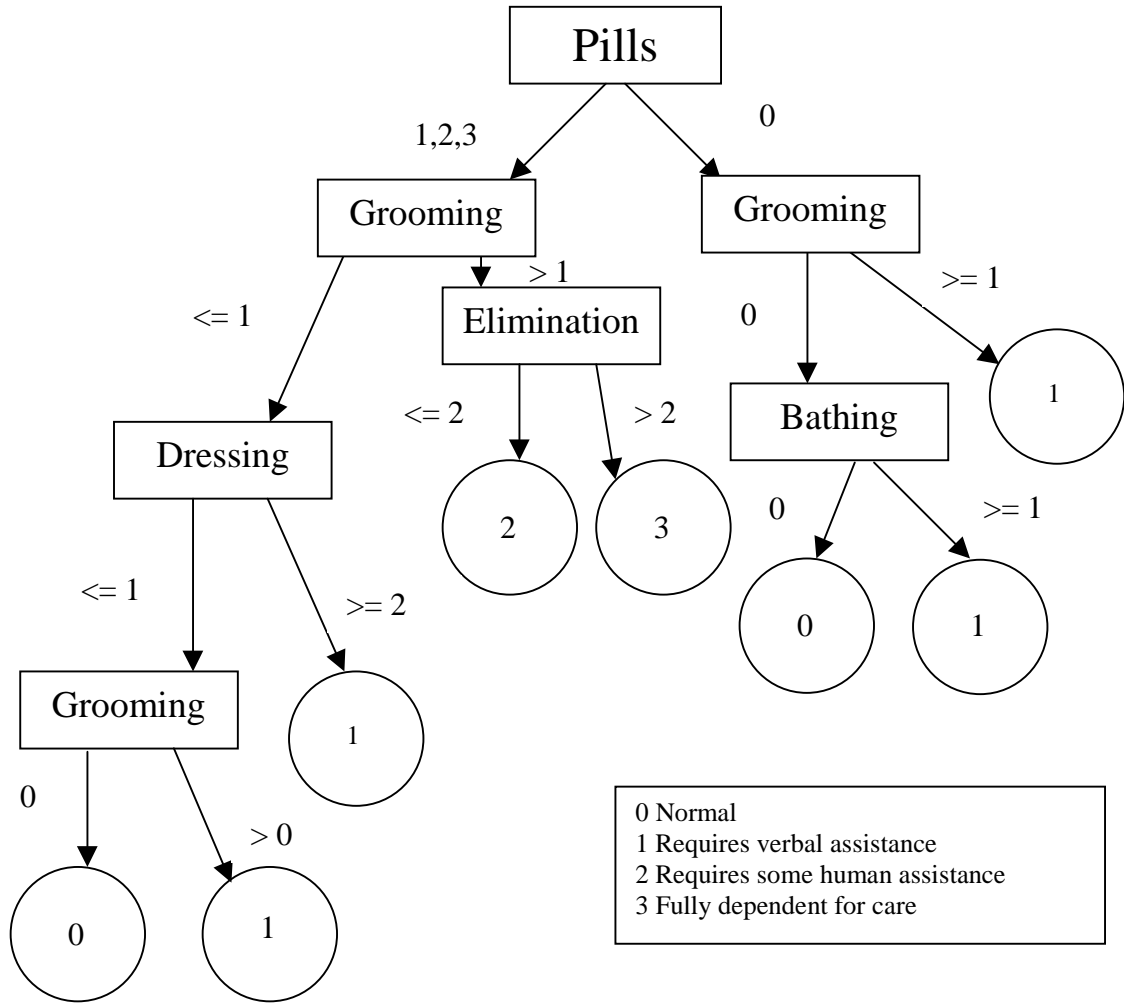
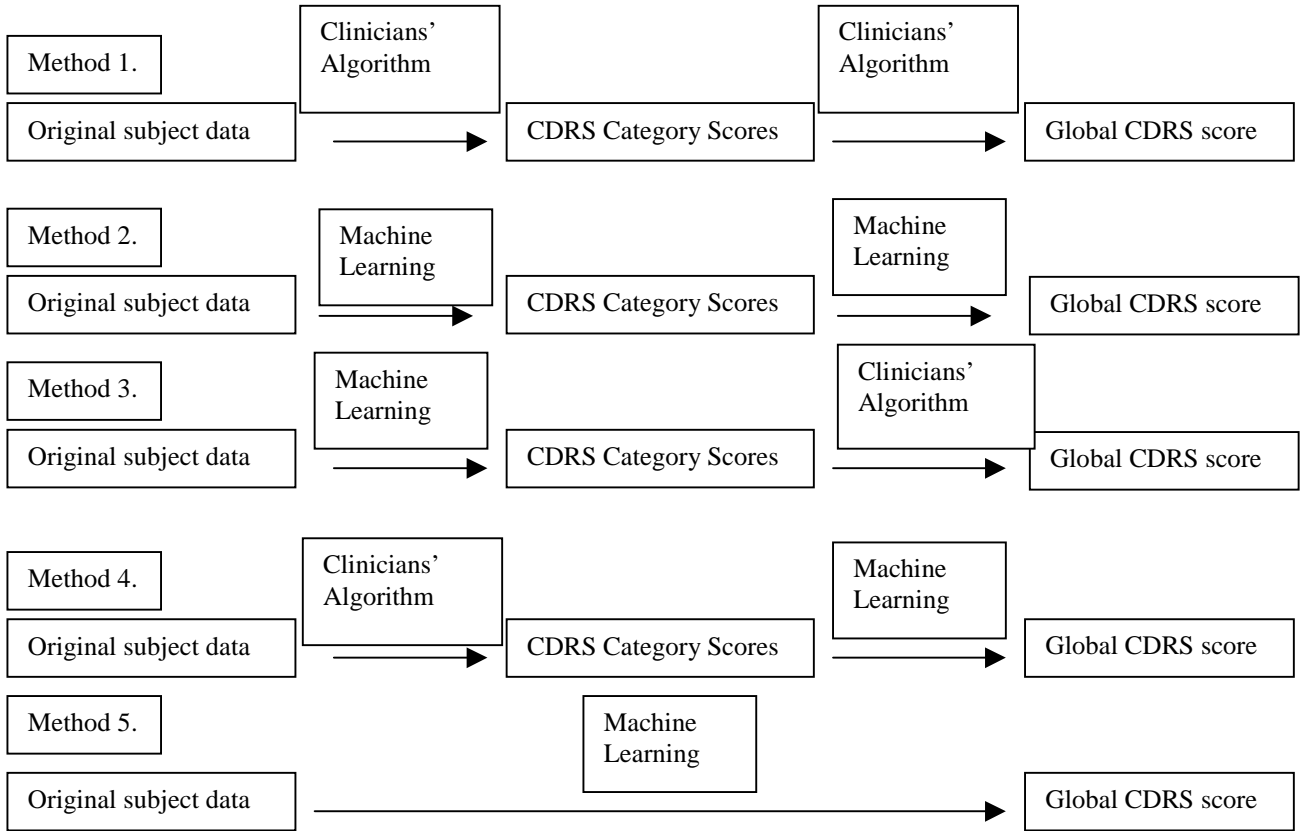


Figure 2. Experimental design to analyze the contributions of Machine learning to global and category CDRS classification.



Effect of Machine Learning both category and global CDRS scores = Accuracy of Method 2 minus accuracy of Method 1. (Method 1 is considered to be the gold standard).

Effect of Machine Learning the global CDRS score directly from the original data vs. using an intermediate learning stage (category CDRS scores) = Accuracy of Method 5 minus accuracy of Method 2.

Effect of Machine Learning the CDRS category scores from the original data = Accuracy of Method 5 minus accuracy of Method 1.

The effect on classification accuracy due to Machine Learning the category CDRS scores = Accuracy of Method 2 minus accuracy of Method 4.

The effect on classification accuracy due to Machine Learning the global CDRS scores = Accuracy of Method 2 minus accuracy of Method 3.

Table 1. Sample breakdown by level of impairment for global CDRS score for 678 subjects and each CDRS category score for the 339 subjects (half the total sample).

CDRS scale (N)	% of sample by CDRS level of impairment				
	0 (normal)	0.5 (very mild)	1 (mild)	2 (moderate)	3 (severe)
Global CDRS (678)	8.8%	28.3%	24.8%	35.4%	2.7%
Memory (339)	15.0%	7.5%	15.5%	30.1%	31.9%
Orientation (339)	20.6%	--	20.9%	53.7%	4.8%
Judg. & Prob. Solv. (339)	18.1%	24.8%	8.1%	37.5%	11.5%
Community Affairs (339)	20.2%	28.3%	30.1%	18.1%	3.3%
Home & Hobbies (339)	28.8%	16.4%	22.9%	24.6%	7.3%
Personal Care (339)	37.9%	--	44.0%	16.1%	2.0%

Table 2. Mean Classification Accuracy (%) of Machine Learning Algorithms for CDRS category and global scores† based on 20 training and testing runs for each algorithm and CDRS scale.

CDR scale	C4.5	C4.5Rules	CART	Naïve Bayes	ADCSU
Global CDRS ⁴	86.3% (2.3)	85.8% (2.3)	82.9% (2.2)	84.1% (2.4)	--
Global CDRS ²	77.0% (4.0)	75.2% (4.1)	70.3% (3.9)	77.9% (3.9)	78.4% ³
Global CDRS ⁵	68.6% (2.9)	63.9% (3.0)	68.4% (3.85)	71.5% (2.9)	
Memory	76.1%	83.2%	78.9%	79.7%	--
Orientation	100.0%	100.0%	99.3%	100.0%	--
Judgment & Problem Solving	69.0%	65.5%	68.5%	65.5%	--
Community Affairs	73.5%	78.8%	77.7%	78.8%	--
Home & Hobbies	61.9%	63.7%	70.7%	81.4%	--
Personal Care	93.8%	96.5%	92.5%	96.5%	--

⁴Global CDRS score was derived using Method 4. ²Global CDRS score was derived using Method 2.

³Global CDRS score was derived using Method 3. ⁵Global CDRS score was derived using Method 5. (See Figure 2)

†Standard Deviation also given in brackets for the Global CDRS scores.

Table 3. Effects of the different methods of applying ML algorithms for CDRS classification accuracy.

Effect (see Figure 2)	C4.5	C4.5Rules	CART	Naïve Bayes	ADCSU
Effect of learning category and global CDRS scores ¹ vs. gold standard	-23%	-24.8%	-29.7%	-22.1%	--
Effect of learning just the category CDRS scores vs. the gold standard ² .	--	--	--	--	-21.6%
Effect of learning global CDRS using an intermediate learning stage vs. directly learning from data ³	8.4%	11.3%	1.9%	6.4%	--
Effect of learning global CDRS without an intermediate stage vs. gold standard ⁴	-31.4%	-36.1%	-31.6%	-28.5%	--
Effect of learning the global CDRS from gold standard category scores vs. Two-Stage learning ⁵	-9.3%	-10.6%	-12.6%	-6.2%	--
Effect of Two-Stage learning vs. using gold standard to derive global CDR from learned category scores ⁶	-1.4%	-3.2%	-8.1%	-0.5%	--

¹Accuracy of Method 2 minus Accuracy of Method 1. ²Accuracy of Method 3 minus Accuracy of Method 1. ³Accuracy of Method 2 minus Accuracy of Method 5.

⁴Accuracy of Method 5 minus Accuracy of Method 1. ⁵Accuracy of Method 2 minus Accuracy of Method 4.

⁶Accuracy of Method 2 minus Accuracy of Method 3.

Table 4. Degree of global CDRS score misclassification by Naïve Bayes using, as input attributes, the clinician-derived and learned (given in bold) CDRS category scores.

True Global CDRS score (level)	# of CDRS levels by which the subjects were misclassified.							
	1		2		1		2	
	Global CDRS Classification Method							
	NB ^a	NB ^b	NB ^a	NB ^b	NB ^a	NB ^b	NB ^a	NB ^b
	% of subjects with the given global CDRS score misclassified as:							
	Less Impaired (global CDRS < true level)				More Impaired (global CDRS > true level)			
0	--	--	--	--	5.9%	0.0%	0.0%	0.0%
0.5	0.0%	8.8%	--	--	18.2%	20.6%	3.0%	0.0%
1	13.5%	17.4%	0.0%	0.0%	9.6%	13.0%	0.0%	0.0%
2	5.9%	20.0%	1.2%	0.0%	3.5%	0.0%	--	--
3	0.0%	0.0%	0.0%	0.0%	--	--	--	--

^aGlobal CDRS was learned by Naïve Bayes from the CDRS category scores derived from the clinicians' computerized scoring algorithm (Method 4, Figure 2).

^bGlobal CDRS was learned by Naïve Bayes from the CDRS category scores, which were also, learned (Method 2, Figure 2).