Refinement of Neuro-psychological tests for dementia screening in a cross cultural population using Machine Learning

Subramani Mani[1,3,5], Malcolm B. Dick[2,3], Michael J. Pazzani[1,3], Evelyn L. Teng[4], Daniel Kempler[4], and I. Maribell Taussig[4]

[1] Department of Information and Computer Science {mani,pazzani}@ics.uci.edu
[2] Department of Neurology mdick@teri.bio.uci.edu
[3] University of California, Irvine
[4] University of Southern California
[5] Center for Biomedical Informatics, University of Pittsburgh

Abstract. This work focused on refining the Cognitive Abilities Screening Instrument (CASI) by selecting a clinically significant subset of tests, and generating simple and useful models for dementia screening in a cross cultural populace. This is a retrospective study of 57 mild-to-moderately demented patients of African-American, Caucasian, Chinese, Hispanic, and Vietnamese origin and an equal number of age matched controls from a cross cultural pool. We used a Knowledge Discovery from Databases (KDD) approach. Decision tree learners (C4.5, CART), rule inducers (C4.5Rules, FOCL) and a reference classifier (Naive Bayes) were the machine learning algorithms used for model building. This study identified a clinically useful subset of CASI, consisting of only twenty Mini Mental State Examination (MMSE) attributes—CASI-MMSE-M, saving test time and cost, while maintaining or improving dementia screening accuracy. Also, the machine learning algorithms (in particular C4.5 and CART) gave stable clinically relevant models for the task of screening with CASI-MMSE-M. ...

1 Introduction

Demand is growing for brief, reliable, and sensitive methods to detect dementia, given the emphasis on cost-effectiveness in the current health care environment. Already, managed care companies are limiting access to traditional neuropsychological testing. As the number of older adults in the United States continues to increase and concern heightens about memory problems, health care professionals will need quick, effective tools to accurately separate cognitively impaired from healthy elders from a variety of cultural backgrounds. The prevalence of dementia has been estimated as ranging from 1-3% in the 65-74 age group, from 7-19% in those 75-84, and from 25-47.2% in those 85 and older [1], [2]. Neuropsychological assessment has retained its key role in the diagnosis of dementia despite improvements in neuroimaging techniques, such as magnetic resonance imaging (MRI) and single photon emission computerized tomography (SPECT).

While forgetfulness has been described as America's latest health obsession [3], dementia continues to be under-recognized within community practice settings. The problem of under-recognition is even greater in minority elders [4] due to a variety of factors. The process of diagnosing dementia in minority individuals is complicated by cultural beliefs about Alzheimer's disease and similar disorders, mistrust of mainstream care providers, the cultural press to "take care of the problem" within the family, economic limitations, and, until recently, a lack of culturally sensitive neuropsychological tools. The Cross-Cultural Neuropsychological Test Battery (CCNB) was developed in response to the growing interest in and need for culturally fair measures of cognitive functioning [5].

2 Methods

2.1 CCNB and CASI

The eleven tests comprising the Cross Cultural Neuropsychological Test Battery (CCNB) assess recent memory, attention, language, reasoning, and visual spatial functioning (areas known to be impaired in Alzheimer's disease) as well as overall mental status. Mental status is assessed with the Cognitive Abilities Screening Instrument (CASI) [6]. Designed for cross-cultural application, the CASI is easier to adapt for a variety of cultural/language groups than many of the screening instruments currently used with English-speaking individuals (e.g., MMSE, BIMC). Unlike these instruments, when direct translation of an item is inappropriate, the CASI provides a culturally fair alternative. For example, as the phrase "No ifs, ands, or buts" is meaningless to non-English speaking individuals, the CASI provides alternative versions of this item, using linguistically equivalent phrases from other languages. Many of the items in the CASI are common to the Mini-Mental State Exam (MMSE) [7], the Modified Mini-Mental State Exam (3MSE) [8], and the Hasegawa Dementia Screening Scale (HDSS) [9]. Consequently, scores on subsets of the CASI are equivalent to scores on the MMSE, 3MSE, and HDSS. In the current health care environment, patients are unlikely to receive neuropsychological testing due to the cost and time involved. Application of the CCNB in clinical settings may be limited by the 90-minute administration time. Consequently, ongoing efforts are directed at finding ways to shorten the CCNB without compromising it's diagnostic accuracy.

3 Description of the data set

The data for this study is from the Los Angeles and Orange County areas of Southern California. Normative data has been collected on a total of 324 healthy English speaking (i.e., Caucasian, African American) and non-English speaking (i.e., Chinese, Vietnamese, Hispanic) minority elders. The investigators are currently expanding the CCNB normative database to other highly represented minority groups in the United States, beginning with Koreans, and administering the battery to cognitively impaired older adults in each of the minority groups,

with a total of 57 demented patients tested to date. Using conventional statistical methods, the investigators [5] demonstrated that most of the tests in the CCNB are culturally fair. Ethnicity had a limited impact on overall test scores, affecting only certain measures (e.g., Digit Span, Category Fluency), while education played a significant role in performance on almost all of the tests. The sample consisted of the 57 mild-to-moderately demented patients of African-American, Caucasian, Chinese, Hispanic, and Vietnamese origin and an equal number of age matched controls from a cross cultural pool. The specific data entered for each subject included gender, age, education, and ethnicity plus responses to all of the items on the CASI. See Table 1 for the sample characteristics.

Table 1. Characteristics of the sample of this study

	Cases (n = 57)		Controls (n = 57)	
Attribute	Mean	Std. Dev.	Mean	Std. Dev.
Age	76.05	9.4	73.93	7.7
Education	10.33	5.2	9.11	4.9

3.1 Neuropsychological Test Data sets

We created four datasets from the whole sample. The CASI-FULL-BATTERY (CASI-FB) has 42 attributes and the outcome variable or class (normal or demented). The CASI-MMSE—CASI-SCORED (CASI-MMSE-C) has a subset of 20 MMSE variables scored in the CASI framework plus the class. The third dataset was the CASI-MMSE—MMSE-SCORED (CASI-MMSE-M) which also had the 20 MMSE variables but scored in the MMSE framework plus the class. We also included a CASI-SHORT (CASI-SH) [6] consisting of just 8 CASI variables plus the class. In this work, we have focused on selecting sets of features corresponding to tests used in screening for dementia. More traditional approaches (e.g., forward and backwards selection procedures [10] consider adding or deleting individual features. These were not considered in this study because it was important to preserve the structure of the existing examinations as much as possible so that there was less resistance to adopting changes in procedures.

3.2 Machine Learning and KDD

Machine Learning (ML) and Knowledge Discovery from Data bases (KDD) are increasingly being applied in health care to build models, develop practice guidelines or refine guidelines for better medical decision making. They differ from traditional approaches by generating domain models such as decision trees, rules, graphs etc. from data. The KDD process involves many steps encompassing data pre-processing (attribute selection, recoding etc.), choice of datamining algorithms, experimental protocol and post-processing of the output. See [11] for a detailed discussion on this. Some recent applications of these techniques in the

medical domain include differential diagnosis of abdominal pain [12], screening and severity staging models for dementia [13], [14] and learning from a database of sports injuries [15]. Using ML and KDD techniques, we are attempting to refine the CCNB with two explicit goals. First, we are interested in identifying a clinically usable subset of CASI (CASI-SUBSET) which will save time and cost retaining or improving the accuracy obtained by the full battery. Second, to generate simple and useful models for dementia screening in a cross cultural population with the CASI-SUBSET, maintaining or improving the sensitivity and specificity obtained using a total score cutoff value. Hopefully, these refinements should lead to greater utilization of the CCNB in clinical settings. As a first attempt at refining the CCNB through ML and KDD techniques, this study examined data from the Cognitive Abilities Screening Instrument (CASI; Teng et al., 1994). The CASI items cover 10 cognitive domains commonly assessed in dementia: attention, concentration, orientation, short-term memory, long-term memory, language ability, constructional praxis, verbal fluency, abstraction, and everyday problem solving skills. In most of these domains, scores range from 0 to 10 points while the total CASI score ranges from 0 to 100. Many of the CASI items are taken directly, or modified, from the Mini Mental State Exam [7]. The subset of CASI items representing the MMSE yields an equivalent score to that achieved on the MMSE itself. This raises the possibility that a shortened version of the CASI, more specifically the subset of items from the MMSE, could achieve the same level of diagnostic accuracy as the entire CASI.

3.3 Machine Learning Algorithms

Of particular interest to this study is the case where guidelines for screening must be written down and followed by an organization. In this case, automated tools that make black-box predictions are not acceptable and the human interpretability of rules and trees is a major benefit. We concentrated on decision tree learners and rule learners as they generate clear descriptions of how the ML method arrives at a particular classification. These models have the advantage that they can easily be taken *offline*, and depicted as charts representing a rule set or decision tree. They also tend to be simple and understandable models, compared with complex models such as neural networks, Bayesian networks or multiple models. Naive Bayes [16] was selected as a reference baseline classifier for comparison purposes.

C4.5 is a decision tree generator and C4.5Rules produces *if ... then* rules from the decision tree [17]. 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. FOCL [18] 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 retarded) and information on each continuous variable on whether higher or lower values of the variable increases the probability of belonging to a class. When this facility of FOCL is used, it is termed "constrained" FOCL. For this study we used only the "unconstrained" functionality of FOCL. CART [19] is a classifier which uses a tree-growing algorithm that minimizes the standard error of the classification accuracy based on a particular tree-growing method applied to a series of training subsamples. We used Buntine and Caruana's implementation of CART, (the "IND" package) [20]. For each training set, CART builds a classification tree where the size of the tree is chosen based on cross-validation accuracy on this training set. The accuracy of the chosen tree is then evaluated on the unseen test set.

3.4 Model Building

We used MLC++ [21] to run these ML algorithms on the 4 datasets. For each of the four datasets (set Section 3.1), the whole sample was divided into 10 random partitions of equal size and a ten-fold cross validation was done. The training and test sets were created from these partitions as follows. For each of these partitions P_i, the test set was P_i and the training set was the whole sample S minus P_i i.e. all the other partitions $P_{j(j \neq i)}$. Models were generated from the training set and evaluated on the unseen test set. The classification accuracies reported are the mean scores obtained with the ten test sets. Note that cross validation evaluates each instance only once and the n which goes into the various statistics such as total accuracy, sensitivity and specificity is the size of the whole sample. This methodology is based on the approach recommended by Salzberg [22].

Table 2. Sensitivity and Specificity of the ML algorithms for dementia screening in a cross cultural population with the CASI-FB and CASI-MMSE-M

(Total sample $n = 114$, Impaired/Demented $n = 57$, Normal $n = 57$)

	CASI-FB			CASI-MMSE-M		
Algorithm	Accuracy[†]	Sensitivity	Specificity	Accuracy[†]	Sensitivity	Specificity
C4.5	82.46(10.9)	75.44	89.47	84.21(13.6)	80.70	87.72
C4.5Rules	81.58(14.2)	75.44	87.72	83.33(15.8)	77.19	89.47
Naive Bayes	85.96(10.8)	77.19	94.74	82.46(10.8)	73.68	91.23
CART	73.88(14.8)	58.77	89.06	79.33(18.9)	69.65	89.06
FOCL	84.20 (5.8)	75.44	92.98	78.10(13.4)	84.21	71.93

[†] The standard deviation is given in braces.

ML—Machine Learning, CASI—Cognitive Abilities Screening Instrument, CASI-FB—CASI full battery, CASI-MMSE-M—MMSE subset of CASI with MMSE scoring scheme

Table 3. Sensitivity and Specificity of the ML algorithms for dementia screening in a cross cultural population with the CASI-MMSE-C and CASI-SH

(Total sample $n = 114$, Impaired/Demented $n = 57$, Normal $n = 57$)

	CASI-MMSE-C			CASI-SH		
Algorithm	Accuracy[†]	Sensitivity	Specificity	Accuracy[†]	Sensitivity	Specificity
C4.5	82.46(12.2)	77.19	87.72	78.95(10.5)	82.46	75.44
C4.5Rules	78.95(15.8)	71.93	85.96	78.95(10.5)	75.44	82.46
Naive Bayes	84.21 (9.6)	75.44	92.98	83.33(10.3)	75.44	91.23
CART	72.97(15.8)	60.58	85.42	75.94(10.1)	76.90	74.97
FOCL	84.20 (8.7)	87.72	80.70	78.90(12.9)	80.70	77.19

[†] The standard deviation is given in braces.

ML—Machine Learning, CASI—Cognitive Abilities Screening Instrument, CASI-MMSE-C—MMSE subset of CASI with CASI scoring scheme, CASI-SH—Short CASI

4 Results

Table 2 and Table 3 give the detailed results obtained with the various ML algorithms using the four different datasets—CASI-FB, CASI-MMSE-C, CASI-MMSE-M and CASI-SH. The performance of each of the different algorithms was comparable, across the four neuropsychological tests—CASI-FB, CASI-MMSE-C, CASI-MMSE-M and CASI-SH. Likewise, for the same neuropsychological test, the various algorithms reported similar accuracy figures. In general the performance using the CASI-SH was somewhat lower. Also, CART gave lower accuracy figures across the data sets. We examined the sensitivity (probability of correctly classifying cases, i.e. the demented group in our sample) and specificity (probability of correctly classifying controls, i.e. the normal group) for the testing samples. All the ML algorithms except FOCL gave higher specificity across CASI-FB, CASI-MMSE-C and CASI-MMSE-M datasets while in the case of CASI-SH sensitivity was higher with C4.5, CART and FOCL. CART gave typically simple decision trees with two to six leaves. See Figure 1 for a representative CART tree. C4.5 gave slightly larger trees with the number of leaves ranging between two and eleven. The rules generated by C4.5Rules were comparatively simpler. See Figure 2 for a typical set of C4.5 rules.

5 Discussion

CASI-MMSE-M dataset gave higher accuracies with C4.5, C4.5Rules and CART while the accuracy was slightly lower with Naive Bayes. With FOCL, accuracy was lower using CASI-MMSE-M but using CASI-MMSE-C, the performance was similar to that of the full CASI. Note that accuracy is a good metric for evaluating the different datasets since our sample had equal number of cases and controls. Hence, the accuracy is the mean of the sensitivity and specificity. Both the datasets CASI-MMSE-C and CASI-MMSE-M contain the same subset of twenty attributes from the CASI-FB, the only difference being in the scoring of

Fig. 1. Cart tree with five leaves

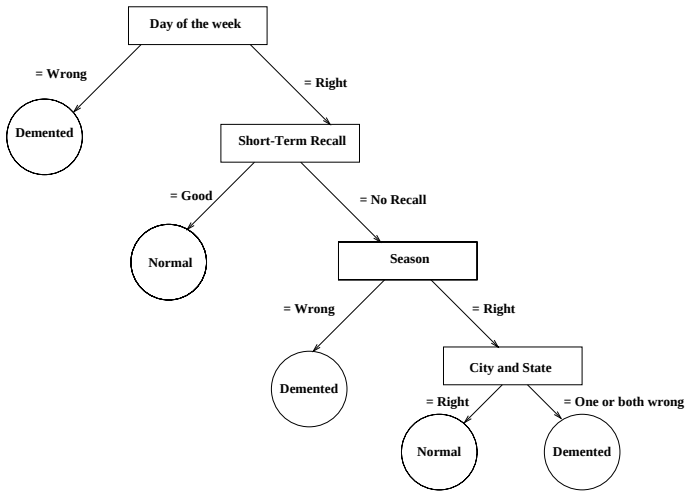

Fig. 2. A C4.5Rule Set

Rule 1: DAY-OF-WEEK = Wrong ⇒ class **demented**(95.8%)
Rule 2: SHORT-TERM-RECALL = Good *and* SEASON = Wrong ⇒ class **demented**(93.0%)
Rule 3: DAY-OF-WEEK = Right ⇒ class **normal**(71.3%)
Default: ⇒ class **demented**

Note: These rules are applied sequentially; the percentage figures in paretheses alongside each rule are the accuracy figures of the rule when it is applicable.

these tests (see Section 3.1). Since MMSE has been used extensively since 1975 and easier to score, the CASI-MMSE-M which makes it culturally fair could be substituted for MMSE in a cross cultural population. It will also save testing time and hence cost when compared to CASI-FB, while improving or retaining test accuracy. CASI-MMSE-M satisfies our *first refinement goal* of identifying a clinically usable subset of CASI saving provider time and cost without compromising on accuracy. In this case actually we improve accuracy marginally. C4.5 gave the highest classification accuracy (84.21%) and sensitivity (80.70%) with the CASI-MMSE-M but the performance of the other algorithms excepting FOCL were comparable. C4.5Rules came close with an accuracy of 83.33% and sensitivity of 77.19%. Decision trees and rule sets are considerably more understandable models when compared for example to Naive Bayes which gives a probability density function. Hence they have the potential to be much more useful in clinical practice.

The basic factors in model selection are its accuracy, comprehensibility and stability, and in medical domains comprehensibility is particularly important. We have addressed these issues in further detail elsewhere [14]. Here we present some interesting properties of our models. In general CART models scored high on comprehensibility compared to C4.5 trees. They were smaller and had fewer domain constraint violations.

5.1 Highlights of our CASI-MMSE-M model

Table 2 (last 3 columns) gives the accuracy, sensitivity and specificity of the various ML algorithms using CASI-MMSE-M. The total accuracy, sensitivity and specificity of the different algorithms are clinically acceptable. The decision tree learners, C4.5 and CART also gave stable models. The attribute *Day-of-week* formed the root of C4.5 trees in nine out ten models and six out of ten CART trees. A CART or C4.5 decision tree with *Day-of-week* as root could serve as a good starting point for selecting a good screening model. The other important consideration apart from accuracy is faithfulness to domain principles and constraints. There is preliminary evidence that domain faithfulness plays a significant role in model acceptability by health care providers [23]. The CART tree in Figure 1 did not violate any domain constraints.

The CART tree selected by the expert (Figure 1) had four attributes while the C4.5 rule set in Figure 2 had just three. Though the CASI-MMSE-M is composed of twenty attributes, most models we generated comprised of four to six features. This clearly shows that all the attributes in the CASI-MMSE-M are not equally significant. Moreover, attributes such as *Day-of-week*, *Short-term-recall* and *Season* figure at the top region of the trees and rule-sets consistently. This raises the possibility of a shorter test compared to CASI-MMSE-M. Further research is required before an optimal subset of features could be advanced as a shorter test. On the other hand, a model such as the one in Figure 1 could be used in community settings for dementia screening in a cross-cultural population. This completes our *second refinement task* of identifying a clinically useful model for dementia screening in a cross cultural population with the CASI-SUBSET, the CASI-MMSE-M.

5.2 Limitations and Conclusions

The dataset which we used for our model building task came from a representative cross cultural population, but the sample size was small ($n = 114$). The study findings would have to be verified with a larger sample before adopting it. Likewise, the proportion we have used (equal number of cases and controls) is not reflective of the general population. Hence the models will have to be validated for the relevant population groups. Furthermore, for the task of dementia screening, the CASI-MMSE-M might turn out to be optimal saving provider time and cost while retaining or improving screening accuracy, but the smaller instrument might be insufficient for the task of dementia staging or differential diagnosis.

In this study, we have shown that ML algorithms can be employed successfully in refining a battery of neuro-psychological tests (CASI) suitable for a cross cultural populace. ML methods achieved two explicit goals. First, this study identified a clinically usable subset of CASI consisting of only twenty attributes (CASI-MMSE-M) saving test time and cost while maintaining or improving dementia screening accuracy. Second, the ML algorithms in particular the decision tree learners C4.5 and CART gave stable clinically usable models for the task of screening with CASI-MMSE-M. The study is also important from the perspective of the use of ML and KDD methods to the novel task of refinement of a neuro-psychological test battery.

Acknowledgements This work was supported in part by the National Library of Medicine postdoc trainee fellowship to SM. We thank the two anonymous reviewers for their helpful comments and suggestions.

References

1. D. A. Evans, H. H. Funkenstein, M. S. Albert, P. A. Scherr, N. R. Cook, M. J. Chown, L. E. Hebert, C. H. Hennekens, and J. O. Taylor. Prevalence of alzheimer's disease in a community population of older persons. *Journal of the American Medical Association*, 262:2551–2556, 1989.
2. Losing a million minds: Confronting the tragedy of alzheimer's disease and other dementias. U.S. Congress, Office of Technology Assessment, Washington D.C., 1987. Publication OTABA-323.
3. G. Cowley and A Underwood. Our latest health obsession: Memory. *Newsweek*, pages 49–54, June 15 1998.
4. G. Yeo, D. Gallegher-Thompson, and M. Lieberman. Variations in dementia characteristics by ethnic category. In G. Yeo and D. Gallagher-Thompson, editors, *Ethicity and the dementias*, pages 21–30. Taylor & Francis, Washington, D.C., 1996.
5. M. B. Dick, E. L. Teng, D. Kempler, D. S. Davis, and I. M. Taussig. The cross-cultural neuropsychological test battery (ccnb): Effects of age, education, and ethnicity on performance. submitted, 1998.
6. E. L. Teng, K. Hasegawa, A. Homma, Y. Imai, A. Larson, E.and Graves, K. Sugimoto, T. Yamaguchi, H. Sasaki, D. Chiu, and L. R. White. The cognitive abilities screening instrument (casi): A practical test for cross-cultural epidemiological studies of dementia. *International Psychogeriatrics*, 6:45–58, 1994.
7. MF Folstein, SE Folstein, and PR 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.
8. E. L. Teng and H. C. Chui. The modified mini-mental state (3ms) examination. *Journal of Clinical Psychiatry*, 48:314–318, 1987.
9. K. Hasewaga. The clinical assessment of dementia in the aged: A dementia screening scale for psychogeriatric patients. In M. Bergener, U. Lehr, E. Lang, and R. Schmitz-Scherzer, editors, *Aging in the eighties and beyond*, pages 207–218. Springer, New York, 1983.
10. R Caruana and D Freitag. Greedy attribute selection. In W Cohen and H Hirsh, editors, *Machine Learning: Proceedings of the Eleventh International Conference*. Morgan Kaufmann, 1994.

11. Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery: An overview. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–36. AAAI Press, Menlo Park, California 94025, 1996.
12. C Ohmann, Q Yang, V Moustakis, K Lang, and van PJ 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, pages 276–281. Springer, 1995.
13. WR Shankle, S Mani, M Pazzani, and P Smyth. Detecting very early stages of dementia from normal aging with machine learning methods. In Elpida Keravnou, Catherine Garbay, Robert Baud, and Jeremy Wyatt, editors, *Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine, AIME97*, volume 1211, pages 73–85. Springer, 1997.
14. Subramani Mani, William R. Shankle, Malcolm B. Dick, and Michael J. Pazzani. Two-Stage Machine Learning Model for Guideline Development. *Artificial Intelligence in Medicine*, 1998. In Press.
15. I.Zelic, I.Kononenko, N.Lavrac, and V.Vuga. Machine learning applied to diagnosis of sport injuries. In Elpida Keravnou, Catherine Garbay, Robert Baud, and Jeremy Wyatt, editors, *Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine, AIME97*, volume 1211, pages 138–144. Springer, 1997.
16. RO Duda and PE Hart. *Pattern Classification and Scene Analysis*. John Wiley, New York, 1973.
17. JR Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, Los Altos, California, 1993.
18. Michael Pazzani and Dennis Kibler. The Utility of Knowledge in Inductive Learning. *Machine Learning*, 9:57–94, 1992.
19. L Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.
20. Wray Buntine and Rich Caruana. *Introduction to IND Version 2.1 and Recursive Partitioning*. NASA, 1992.
21. R Kohavi, George John, Richard Long, David Manley, and Karl Pfleger. MLC++: A machine learning library in C++. In *Tools with Artificial Intelligence*, pages 740–743. IEEE Computer Society Press, 1994.
22. Steven L. Salzberg. On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. *Data Mining and Knowledge Discovery*, 1:317–328, 1997.
23. Michael J. Pazzani, Subramani Mani, and W.R. Shankle. Beyond concise and colorful: Learning intelligible rules. In *The third international conference on Knowledge Discovery and Datamining*, pages 235–238. AAAI Press, Menlo Park, California., 1997.