Fuzzy ARTMAP Neural Network Compared to Linear Discriminant Analysis Prediction of the Length of Hospital Stay in Patients with Pneumonia

Philip H. Goodman, Vassilis G. Kaburlasos, Dwight D. Egbert
Departments of Medicine and Electrical Engineering, University of Nevada
Reno, NV 89557

Gail A. Carpenter, Stephen Grossberg, John H. Reynolds, David B. Rosen
Department of Cognitive and Neural Systems, Boston University
Boston, MA 02215

Arthur J. Hartz
Division of Biostatistics/Clinical Epidemiology, Medical College of Wisconsin
Milwaukee, WI 53226

Abstract **Health care databases may comprise hundreds of predictive variables, thousands of cases, and complex outcomes. Artificial neural networks may provide an alternative to established predictive algorithms for analyzing massive health care databases, potentially overcoming obstacles arising from the number of cases, missing data, variable selection, multicollinearity, specification of important interactions, and sensitivity to erroneous values. On an actual database derived from patients hospitalized with pneumonia, we compared the cross-validated predictions of linear discriminant analysis (LDA) to a new, supervised adaptive resonance theory network called ARTMAP. Unbiased proportionate reduction in error using ARTMAP was 50% greater than LDA. Under conditions of simulated noise and increasing-proportion learning, ARTMAP demonstrated further advantages over LDA. The promising performance of ARTMAP warrants further evaluation on larger health care databases.**

I. INTRODUCTION

A major national effort is underway to determine patterns of medical practice that most effectively result in favorable health outcomes [1], [2]. Databases arising from medical effectiveness research may contain tens of thousands of cases and hundreds of variables intended to predict outcome status. Established statistical prediction algorithms may be suboptimal for such tasks because of obstacles arising from massive number of cases, missing data, variable selection, multicollinearity, specification of important interactions, and sensitivity to erroneous values. Artificial neural networks may offer an alternative method that overcomes some of the aforementioned analytic problems.

To address these inadequacies, we developed a new self-organizing supervised neural network that incorporates fuzzy set logic into adaptive resonance theory mapping (ARTMAP) to simultaneously predict outcome and define category patterns within outcomes. In voting fuzzy ARTMAP, multiple subnetworks resulting from random permutations of a learning set are created until a stable "voting" consensus (predictive score) is achieved.

The purpose of this study was to determine whether such a self-organizing neural network could accurately predict the length of stay of patients admitted to a community hospital with a diagnosis of pneumonia. Comparison was made to the performance of linear discriminant analysis, the most suitable of the established predictive methodologies.

II. METHODS

A. Clinical Database

1) Predictive Measures: The database was generated by one of us (AJH) through abstraction of 239 charts carrying a principal diagnosis of pneumonia on patients hospitalized between 1988 and 1990 at the Medical College of Wisconsin (MCW) affiliated hospitals. Stepwise linear regression, performed on about 200 clinically tenable measurements, resulted in 16 factors documented within the first 2 days of admission on the 214 patients whose charts had no missing data. The Reno and Boston researchers were blinded to the relative importance of the 16 variables. Risk factors included continuous as well as binary measurements (displayed as part of Fig. 5).

2) Outcome Measure: The outcome measure chosen for this preliminary study was length of stay (LOS), because it is related to both the severity of illness and the process of care, and is a major determinant of the cost of medical care. We broke the LOS into 3 intervals, based on inspection of its distribution (Fig. 1). The distribution is skewed rightward, with a mean LOS of 6.9 days and a median LOS of 5 days.

0-7803-0720-8/92 $3.00

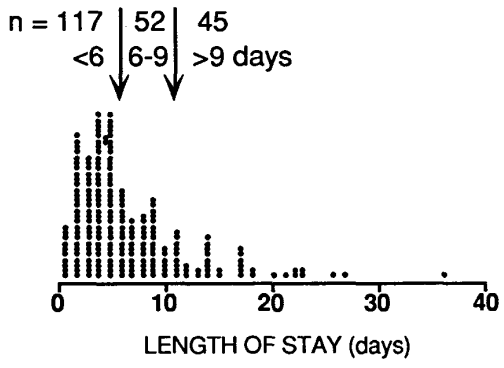

Fig. 1. Distribution of the length of stay for 214 patients hospitalized with pneumonia. Arrows indicate empiric cut-points for trichotomous classification of the length of stay.

Short LOS reflected either relatively healthy patients who responded briskly to therapy, or those who were very ill and died despite treatment. Long LOS reflected relatively sick patients, often requiring intensive care, who usually survived after prolonged hospitalization. There were only 10 deaths, distributed across all 3 LOS categories.

B. Accuracy, Cross-Validation, and Simulations

1) Accuracy: In order to apply algorithmic prediction in a clinical setting, cut-points or thresholds must be used to categorized the predicted outcomes. An ideal predictive algorithm would create probabilities clustered near 0 and 1 for each predicted outcome, so that the accuracy (as judged by the fraction of true positives and negatives) would be insensitive to the choice of a particular cut-point. For both fuzzy ARTMAP and linear discriminant analysis (LDA), we used the predicted outcome with the greatest probability. In the case of LDA, the predictive formula was adjusted for the prior probabilities of each length of stay category. In order to adjust accuracy for chance correctness of classification, we made use of the proportionate reduction in error (PRE) statistic [3],

$$\text{PRE} = \frac{\text{(total correct} - \text{expected correct by chance)}}{\text{(total number of cases} - \text{expected correct by chance)}}$$

The PRE is zero, therefore, when the diagonal sum in a confusion matrix is equal to that expected by chance alone, and 1 when classification is perfect.

2) Cross-Validation: In the analysis of large datasets, validity is threatened more by bias than variance. If a dataset in trained then tested on the same cases (i.e., resubstituted), the predictive accuracy is favorably biased [4]. We obtained nearly unbiased estimates for our dataset by the appropriate use of the k-fold cross-validation technique. In k-fold cross-validation, the data set is randomly divided into k partitions of approximately equal size; the cases not found in each partition are used to train the classifier, and the partitioned cases are used as a testing set. This is performed on all k partitions and the overall predicted class assignments are tallied, from which average accuracy is computed. We varied k from 2 to 100, and found no substantial loss of accuracy with k=10. This is supported by empirical evidence that the partition fraction can approximate the prevalence of each class without significant loss of accuracy [4].

3) Simulations: In the first simulation, we assessed the robustness of the techniques to uncorrelated noise by adding 16 uniform random noise variables to the dataset of 16 existing pneumonia variables, for a total of 32 predictive variables. In the second simulation, we created a larger but correlated data set by replicating each patient record 3 times, distorting the continuous variables by a random positive or negative amount (within one standard deviation). On 10 random permutations of this new 642-record database, we then performed an increasing-proportion comparison of the cross-validated predictive abilities of fuzzy ARTMAP and LDA, using a progressively larger portion of a randomly permuted data set for training, and the remainder for testing.

C. Linear Discriminant Analysis (LDA)

Most of the published studies on predicting outcome from severity-of-illness and treatment utilized linear regression, analysis of variance, LDA, or logistic regression. Each of these models makes distributional assumptions. For instance, dichotomous outcomes like mortality are better modeled by logistic regression if the predictor variables are binary or not normally distributed, whereas LDA performs better if normality holds [5], [6]. Continuous measurements (e.g., LOS) can be broken into intervals for analysis of variance or discriminant analysis, or predicted directly with linear regression. Because LOS was broken into 3 levels, we employed LDA. Analysis was performed using SYSTAT version 5 on an Apple Macintosh platform.

D. Fuzzy ARTMAP

Adaptive resonance theory (ART) neural networks use feedback and competition (analogous to interneuronal and recurrent neural circuits) to self-organize stable recognition codes in real time in response to arbitrary sequences of input patterns. Within the ART architecture, the process of adaptive pattern recognition is a special case of the more general cognitive process of hypothesis discovery, testing,

search, classification, and learning. This property opens up the possibility of applying ART systems to the more general problem of adaptively processing large abstract information sources and databases. The development stems from the formulation of synaptic learning as compartmental interactions characterized by differential equations. Fortunately, competitive models can be formulated as Liapunov functions, so the system is asymptotically stable. Since a parallel architecture described by differential equations can be modeled on a von Neumann computer, we can experiment with neural networks on a multipurpose computer and reserve parallel hardware development for specific applications.

The original ART paradigm of Carpenter and Grossberg [7], called ART1, clustered only binary variables. Subsequently, Carpenter and Grossberg developed ART2 [8] and ART3 [9] for analog variables wherein the similarity of new input vectors to existing category vectors was determined Euclidean dot-products. This scheme required the addition of substantial circuitry to automatically scale input vectors. Recently, these Boston University Center for Adaptive Systems researchers proposed an alternative way to represent nonbinary variables using fuzzy set membership theory [10]. Minimal modification of the ART1 was required, as nonbinary variables could be normalized to the 0-1 range, and a fuzzy subsethood membership function [11] substituted for the ART1 matching formula (which reduces to ART1 if the vectors contain only binary elements). The key difference is that the choice and vigilance equations use the logical "AND" function in ART1 but the "MIN" in fuzzy ART. For example, if a newly input vector (1,1,0,1) is being tested for degree of match to an existing pattern vector (1,0,1,1), the AND operator results in the vector (1,0,0,1). The MIN operator selects the minimum of the 2 values for each variable in the vector, which would result in the same vector as the AND function when only binary data is used. For the case of analog data, consider the vector (1, .8, .2, .7) being tested for goodness-of-category match with the existing pattern (1,0,1,1); the MIN operator produces the vector (1, 0, .2, .7). Both the choice function, T_j (j is the index for the F2 nodes), and the vigilance, ρ, use the L1 norm (absolute sum) of this resultant vector. For either ART1 or fuzzy ART, T_j is maximal when the intersection of a newly input F1 vector with an F2 category vector is identical to the F2 vector norm. To deal with ties, the parameter a exerts a normalizing effect using bottom-up weights (w_{ij}), so that T_j will be maximal for the F2 category with the greatest absolute norm.

To enable ART to learn from experience, or map from input to outcome data vectors, a supervised architecture called adaptive resonance theory mapping, or ARTMAP, was proposed by the Boston researchers, first incorporating binary vectors [12], and later generalized to analog vectors using fuzzy ART [13]. As shown in Fig. 2, ARTMAP utilizes 2 ART modules, one to cluster input variables (ART_a) and one to cluster outcome variables (ART_b), with linkage by a map field of nodes (F_{ab}). While ART_b operates like a typical ART module, clustering multiple patterns of outcome if necessary before presentation to F_{ab}, ART_a function is modified to predictively optimize the formation of its F2 category patterns (which have mutually exclusive assignments, or expectations, for outcome classification). This is accomplished during training by elevating ρ_a on a vector-by-vector basis until the best F2 choice (T_j maximum) for that vector is assigned to the outcome class expected by ART_b. Upon successful assignment, the winning F2 category pattern is modified, or updated, to reflect the impact of the new vector. If the vector is allowed to have maximal impact, the learning is called "fast". Slow learning simply takes a weighted average of the new vector (**b**) with the pre-existing category vector (**1–b**). Training and prediction can occur in real time, since ARTMAP makes an outcome prediction upon presentation of each new input vector; only if it learns the correctness of the prediction does it change the F2 coding, which will affect the subsequent input vector's prediction, and so on. If the data records are presented in a relatively unbiased

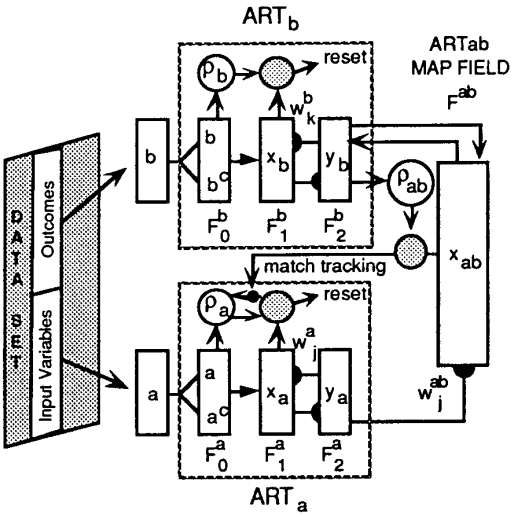

Fig. 2. ARTMAP Architecture (Supervised Learning).

TRAINING: Measurement vectors (a) are preprocessed into (a,a^c) by 0-1 normalization and creation of a complement for each variable. Likewise, (b) are preprocessed into (b,b^c). (a,a^c) is distributed by F1 to all existing F2 category nodes, which feed back learned category weights to F1. The outcome class assignment of the best F2 category match meeting the current vigilance (ρ_a) setting is sent to Fab; if the prediction is confirmed, the F2 category is modified as it learns (a,a^c). If the prediction by ART_a is disconfirmed, MAP FIELD activation induces the match tracking process, raising ρ_a just above the F_1 to F_0 match ratio $|x_a|/|(a,a^c)|$. This triggers another ART_a search for the next best F2 category match, and so on, which leads ultimately to match with an existing ART_a category pattern that correctly predicts (b), or, if none exist, to the assignment of pattern (a) to a previously uncommitted ART_a F2 node. In voting ARTMAP, permutations of (a,a^c) are simultaneously processed in multiple ARTa strata, which send predictions into ARTab, where the consensus is judged.

TESTING: Each new (a,a^c) input to the trained ARTa finds its best F2 match; the outcome class assigned to that F2 category node is the prediction. In voting ARTMAP, the consensus of multiple ARTa strata is used as the final prediction.

fashion, as would be the case in the real temporal sequence of hospital admissions (which we simulate by random permutation), then most of the learning occurs in a single pass through the data, i.e., an ARTMAP program could be used on-line as a real-time "self-taught" expert system. However, a modest increase in accuracy (several percent) is achieved by allowing the input vectors to cycle through several times, until there are no "re-learning" effects on the early-formed F2 categories due to late case presentation. In our experience, only a few cycles, or epochs, of training occur until all input training vectors are learned 100% correctly. This is in contrast to the backpropagation model, which typically must cycle through a data set thousands of times before converging.

In consultation with the Boston Center, the Reno group fully implemented voting fuzzy ARTMAP using C language during the summer of 1991. The Reno ARTMAP was compiled and run on DOS, Apple Macintosh, and UNIX platforms, including the Cray YMP/2 platform at the University of Nevada National Supercomputing Center for Energy and the Environment.

III. RESULTS

A. Accuracy

Under resubstitution, LDA correctly categorized LOS in 67% of cases when trained and tested on the same data set, whereas fuzzy ARTMAP learned to completely discriminate 100% of cases (Table I). Under 10-fold cross-validation, LDA correctly predicted the LOS category in 52% (PRE 0.20), a single ARTMAP network correctly predicted in 56% of cases (PRE 0.26), and a consensus of 30 voting fuzzy ARTMAP networks correctly predicted in 59% of cases (PRE of 0.30) (Table I).

TABLE I
PREDICTIVE ACCURACY OF MODELS

METHOD	Resubstitution	10-fold Cross-Validation	
	Accuracy	Accuracy	PRE[b]
Linear Discriminant	0.67	0.52	0.20
Single ARTMAP	1.00	0.53	0.22
30-Vote ARTMAP	1.00	0.59	0.30

[a]Resubstitution refers to testing after training on all cases.
[b]PRE is the proportionate reduction in error (see text for explanation).

B. Simulations

With the addition of 16 predictive variables consisting of uncorrelated uniform random noise, the PRE of voting ARTMAP actually improved from 0.257 to 0.265 (a 3% increase), whereas the PRE of LDA decreased from 0.203 to 0.164 (19% decrease) on the same data set.

Under the increasing-proportion simulation (Fig. 3), training on only the first 42 cases resulted in about 55% accuracy (PRE 0.33) on the remaining 600 cases for both ARTMAP and LDA. However, as an increasing portion of the database was used to train (with the remainder used as the independent testing set), the predictive accuracy increased only marginally for LDA to about 62% (the covariance matrix changes minimally), but ARTMAP accuracy increased rapidly to an asymptote at about 90% (PRE 85%) with about 300 records or fewer used for testing. Training with 442 and testing on 200 cases increased the PRE for LDA almost two-fold, but more than tripled the PRE for ARTMAP. These findings demonstrate the superior capacity of ARTMAP over LDA to learn from data regionally clustered in data space (i.e., to recognize similar patterns under noisy or systematically biased conditions).

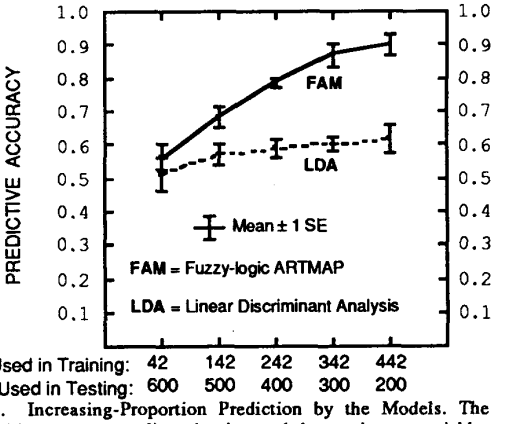

Figure 3. Increasing-Proportion Prediction by the Models. The original 214 cases were replicated twice, and the continuous variables were randomly varied by up to 1 standard deviation. FAM and LDA were trained and tested on an increasing proportion of the new 642-case dataset. The process was repeated 10 times to establish mean accuracies and standard errors.

C. Pattern Identification

One reason to prefer ARTMAP or other locally clustering algorithms over global activation models (such as backpropagation) for the analysis of health care data is to be able to "explain" the partitioning of input feature-space. As an example, during resubstitution trial with a single network, ARTMAP created a total of 44 long-term memory patterns for the 214 MCW pneumonia patients, representing a 5-fold reduction in data categories. The sequence of formation and distribution of patients' records by F2 memory pattern is shown in Figure 4. Figure 4 demonstrates that populous, more "generalized" memory patterns tend to be recruited early in the neural network training process, consistent with real-time learning capability. The reason for this is that early memory patterns are those most subject to change by subsequently presented cases; later patterns emerge to classify the more atypical case clusters.

Fig. 5 displays the structure of the 9 memory patterns from Figure 4 that each clustered at least 10 patients. Among the analog features, there is a trend towards longer hospitalization with increasing age, lower hemoglobin (i.e.,

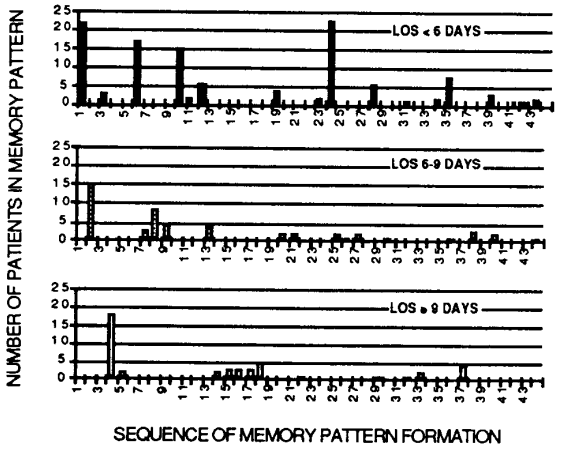

Figure 4.

worsening anemia), and greater time spent in an intensive care unit. Among binary features, short stays demonstrate consistent absence of some risk factors (open rectangles) and inconsistent mixture of others (shaded rectangles). Intermediate stays also cluster the absence of many risk factors, as well as non-white racial background, those without insurance, and arrival by ambulance (filled rectangles). Long stays include categories clustering the blind, those not able to control their urine, not able to walk, and with consistently abnormal laboratory results.

ARTMAP's clustering within outcomes is clinically tenable. For instance, increasing age, worsening anemia, and the presence of risk factors associated with concurrent diseases would reasonably explain, in part, why such patients would require more prolonged hospital treatment. It is conceivable that displays similar to Fig. 5 could enable ARTMAP to "explain itself", alleviating the concern of many physicians and hospital administrators about the use of "black box" predictive algorithms.

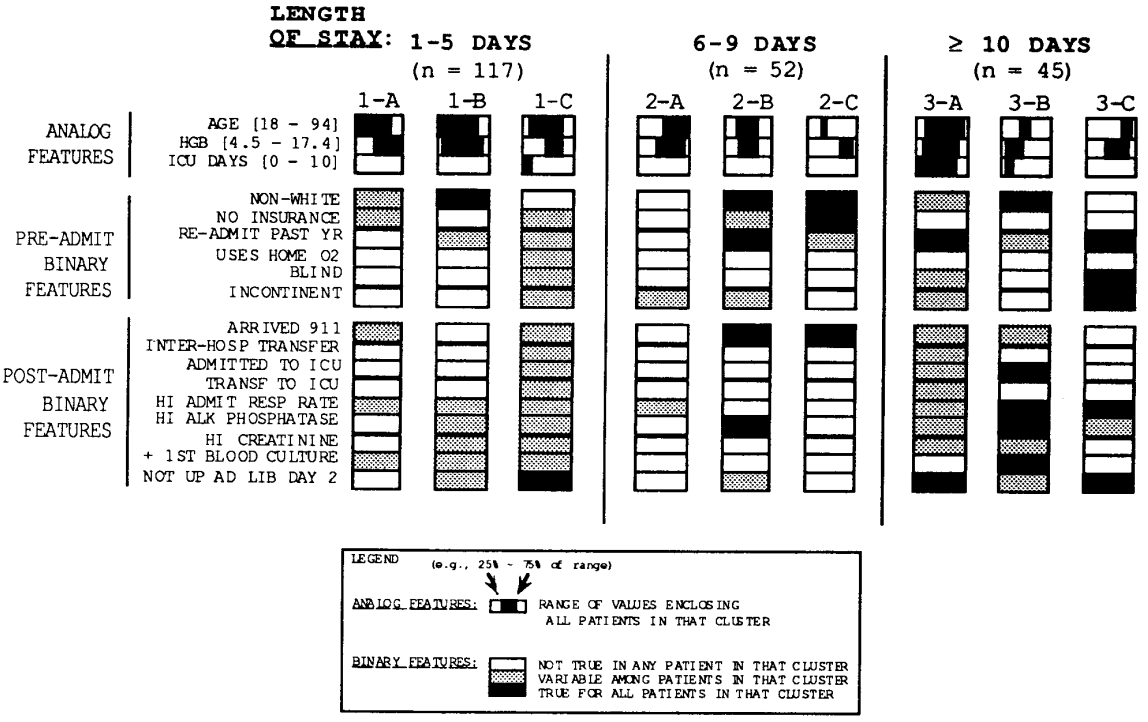

Figure 5. ARTMAP Internal Coding of Categories within Length of Stay Predictions. Explanation of predictive variables: AGE, age in years; HGB, hemoglobin, gm/dl; ICU DAYS, days spent in an intensive care unit; NON-WHITE, race other than white; NO INSURANCE, no private or public source of health insurance; RE-ADMIT PAST YR, admitted at least once to hospital for any problem within the preceding year; USES HOME O2, used continuous home oxygen prior to admission; BLIND, legally blind; INCONTINENT, incontinent of urine; ARRIVED 911, arrived by ambulance; INTER-HOSP TRANSFER, transferred from another hospital; TRANSF to ICU, transferred into an ICU during hospital stay; HI ADMIT RESP RATE, admission respiratory rate > 24; HI ALK PHOSPHATASE, elevated serum alkaline phosphatase; HI CREATININE, serum creatinine > 2; + 1ST BLOOD CULTURE, growth on blood culture from admission; NOT UP AD LIB DAY 2, not able to walk spontaneously by second hospital day.

IV. CONCLUSIONS

The fuzzy ARTMAP consensus of 30 parallel voting networks outperformed the linear discriminant function in predicting length of stay in 214 patients hospitalized for pneumonia (Table I). On this data set, PRE by voting ARTMAP was 50% greater than by LDA (0.30 vs. 0.20). Each ARTMAP layer reached steady state after only 5 to 10 cycles through the dataset.

While the addition of uncorrelated noise variables degraded the predictive function of LDA, it actually improved the accuracy of ARTMAP. It is possible that the ARTMAP system learned to recognize and ignore noise, but this behavior needs to be replicated and further characterized.

Each voting fuzzy ARTMAP network generated outcome-specific multivariate memory categories distinguished by simultaneous ranges within variables (Fig. 5). Memory categories capturing many (populous) and few (sparse) input vectors may both be clinically important. Populous patterns reflect consistent associations among input variables (in our setting, these are severity-of-illness groupings), and thereby contribute to good generalization on future input patterns not previously encountered. In addition, these populous patterns facilitate an "explanation" by the network of those interactions accounting for the predictions. On the other hand, sparse patterns reflect multivariate outliers, resulting from either statistical variation, systematic error, aberrant care (e.g., by hospitals or individual practitioners), or emerging trends (e.g., unrecognized adverse events or new disease entities). In the ARTMAP algorithm, these sparse, outlying memory patterns are not degraded by continued training, reflecting the local nature of the learning paradigm. In our study, inspection of the internal memory patterns revealed that ARTMAP clustered clinically meaningful interactions among the severity-of-illness variables. We are presently exploring the partitioning of variable space by the multiple voting networks in order to improve the "self-explanatory" ability of the system.

Further work should involve larger datasets and missing data. Consideration should also be given to improving predictive performance by tandem or hybrid networks of competitive, locally clustering models like ARTMAP with global activation models like backpropagation.

ACKNOWLEDGMENTS

The authors thank Bahram Nassersharif, Michael Ekedahl, and Sam West of the University of Nevada National Supercomputing Center for Energy and the Environment for their assistance in compiling our software and providing computer time. Supported in part by Washoe Health System, DARPA (AFOSR 90-0083), the National Science Foundation (NSF IRI 90-00530), and the Office of Naval Research (ONR N00014-91-4-4100).

REFERENCES

[1] Goodman PH. The Agency for Health Care Policy and Research: opportunities for research and guidelines leadership. SGIM Newsletter 1990; 13(4):4.

[2] Institute of Medicine. Effectiveness and outcomes in health care. Washington, DC: National Academy Press, 1990.

[3] Liebentrau AM. Measures of association. Beverly Hills: Sage Publ., 1983:16-30.

[4] Weiss SM Kulikowski CA. Computer systems that learn. San Mateo, CA: Morgan Kaufmann, 1991:108-110.

[5] Kandel A. Fuzzy mathematical techniques with applications. Reading, MA: Addison-Wesley, 1986:17-18.

[6] Efron B. The efficiency of logistic regression compared to normal discriminant analysis. J Amer Statistical Assoc 1975;70:892-98.

[7] Carpenter GA, Grossberg S. A massively parallel architecture for a self-organizing neural pattern recognition machine. Computer Vision, Graphics, and Image Processing 1987;37:54-115.

[8] Carpenter GA, Grossberg S. ART 2: self-organization of stable category recognition codes for analog input patterns. Applied Optics 1987;26(23):4919-30.

[9] Carpenter GA, Grossberg S. ART 3: hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. Neural Networks 1990;3:129-52.

[10] Carpenter GA, Grossberg S, Rosen DB. Fuzzy ART: an adaptive resonance algorithm for rapid, stable classification of analog patterns. Neural Networks 1991;6:759-71.

[11] Kosko B. Fuzzy entropy and conditioning. Information Sciences 1986; 40:165-74.

[12] Carpenter GA, Grossberg S, Reynolds JH. ARTMAP: supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural Networks 1991;4:503-44.

[13] Carpenter GA, Grossberg S, Markuzon N, Reynolds JH, Rosen DB. Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Transactions on Neural Networks, in press, 1992.