A Comparison of Three Techniques For Rapid Model Development: An Application in Patient Risk-Stratification

Eric L. Eisenstein[1], DBA and Farrokh Alemi[2], PhD

[1]Outcomes Research And Assessment Group, Duke Clinical Research Institute, Duke University Medical Center, Durham, NC
[2]Health Care Administration Program, Cleveland State University, Cleveland, OH

Accurately risk-stratifying patients is a key component of health care outcomes assessment. And, many health care organizations increasingly are relying upon automated means for assistance in making patient risk-stratification decisions. Unfortunately, the process of outcome model development, as it is currently practiced, is both time consuming and difficult. We investigated the relative abilities of three modeling techniques (logistic regression, artificial neural network (ANN), and Bayesian) to rapidly develop models for risk-stratifying patients. Our results demonstrated that all three modeling techniques perform equally well in certain situations. However, the Bayesian model with conditional independence had the best overall performance. Unfortunately, none of the models were able to achieve the degree of accuracy which would be required in a medical setting.

INTRODUCTION

Risk-stratification is essentially a problem of accurately classifying patients into different risk categories. Whether the health care practitioner is triaging myocardial infarction patients in a chest pain unit or selecting patients with severe coronary artery disease, the problem reduces to one of assigning the patient to the proper category. Classification has been defined as, "a problem solving method that associates a set of categories with possible problem solutions.[1]" In making a classification, the decision maker first evaluates the characteristics of an object and then uses this evaluation to assign the object to one category within a set of possible solutions. In risk-stratification, the patient is the object and the different risk strata are the categories.

Logistic regression has traditionally been used to model medical classification problems. However, recent studies have reported levels of accuracy with ANNs that are greater than those previously reported with logistic regression.[2, 3, 4] Despite these successes, a major impediment to the widespread acceptance of ANNs in medical classification is the fact that they do not provide explanations for their forecasts which are easily interpreted by human operators. Thus, in choosing between ANNs and logistic regression for medical classification modeling there is an implicit choice between forecasting accuracy and a technique whose prescriptions are readily understood. Unfortunately, both of these factors are often important in medical classification problems. Thus, a method for modeling medical classification problems which could equal or exceed the accuracy of ANNs while also providing the intelligible explanations of logistic regression would represent an advance over both of these classification methods.

Bayesian models have been used to model medical classification problems and their accuracy has exceeded that of logistic regression.[5] Bayesian models also have an advantage over ANNs in that their forecasts are based upon likelihood ratios which are easily interpretable by health care professionals. To date, no studies have compared the relative accuracies of all three modeling techniques (logistic regression, ANN, and Bayesian) in the same problem domain. A few medical classification studies have compared Bayesian models with classical statistical techniques or Bayesian models with ANNs.[6,7,8] However, these studies have typically used national averages or expert judgments to create their Bayesian models while they trained the classical statistical techniques and the ANNs on random samples of the same populations they later used for comparative testing. Thus, there was no way of determining whether these other techniques were in fact superior to Bayesian probability models or whether they merely trained on more appropriate data. Recently, researchers have begun to separate the subjective assessment of probabilities from the dynamics of Bayesian probability models.[5,9] This innovation has led to the development of Bayesian learning systems

0195—4210/96/$5.00

which are able to create their own models directly from training data sets.

Chard first tested the relationships between Bayesian model accuracy, training sample size and training attribute set size.[10] His work identified key training sample and attribute set size thresholds which were used as a guide in this study's research design.

MATERIALS AND METHODS

The data in this study were collected under a Health Care Finance Administration grant.[11] Sixteen questions in that data base corresponded to the factors used in the APACHE II system to predict patient outcomes.[12] Another question in this database described patient discharge status (coded as dead or alive) and was used as the outcome which this study's models predicted.

There were a total of 1139 cases in the myocardial infarction database. Initial training samples were selected in sizes of 100 and 600 cases. These sizes were chosen because they were outside the range (200 to 500 cases) previously used by Chard to test the simple Bayes model's performance.[10] Chard reported that there were not enough cases for accurate prediction below his range and that model accuracy did not improve above his range. The test sample size was set at 500 cases. Training and test samples were constructed by selecting the last 500 cases that were enrolled in the database as the testing sample and then randomly selecting two training samples of 100 and 600 cases from the remaining 639 cases, with replacement. This selection method allowed us to simulate a prospective sampling procedure. Two sets of training and testing data were constructed: one in which missing values were allowed and the other in which missing values were imputed as normal.

Attributes were selected in sizes of 3 and 11 APACHE II questions. These sizes were chosen because they were below and above the seven attribute set size previously used in Chard's study.[10] Attributes were selected from the sixteen APACHE II questions ordered by their relative informativeness in predicting mortality for myocardial infarction patients admitted to a hospital . This assessment of relative informativeness was made by a physician managing a hospital intensive care unit.

A total of eight training sample situations were created for our evaluations; four with missing data values imputed as normal and four where missing data values were not imputed. The four training sample situations in each of these groups were: (1) 100 training cases and 3 attributes, (2) 100 training cases and 11 attributes, (3) 600 training cases and 3 attributes, and (4) 600 training cases and 11 attributes.

Three classification techniques were assessed: logistic regression, Bayesian, and ANN. Each classification technique included models with and without interactions. Thus, a total of six individual models were evaluated in this study. The two logistic regression models were called: Logistic Simple and Logistic Interaction. The Logistic Simple model omitted interaction terms and the Logistic Interaction model included them. The two Bayesian models were called: Bayes Simple and Bayes Proper. Since Bayesian systems model attribute interactions through conditional dependence, the Bayes Simple model (without conditional dependence) did not model interactions and the Bayes Proper model (with conditional dependence) modeled interactions. The two ANN models were called ANN 0 Hidden Layers and ANN 1 Hidden Layer. Since ANNs model interactions through hidden network layers, the ANN model with 0 Hidden Layers did not model interactions while the ANN with 1 Hidden Layer did model interactions. The number of ANN hidden layers has been varied in previous studies.[13, 14]

Both logistic regression models were created using SAS's LOGIST procedure in stepwise mode. And, interactions were modeled as separate variables.

Bayesian models were created using a system developed for a previous study.[5] This system enters attributes into Bayesian models according to an a priori assessment of their potential information value. In this assessment, the attribute with the most extreme likelihood ratio among those remaining in the data set is selected next for model inclusion. Two orders of conditioning were used in this study to model attribute interactions. The Bayes Simple model had zero orders of conditioning and the Bayes Proper model implemented conditioning on the two previous attributes entered into the model. Attributes were considered conditionally dependent if the ratio of an attribute's posterior likelihood ratio (after selection) to its prior likelihood ratio (before selection) exceeded 1.2. Thus, conditional dependence measured the change in the prior likelihood ratio that occurred when an attribute was considered for inclusion in the model. The Bayes Simple model, with no conditioning, included all cases in its estimates for each attribute value. Since the Bayes Proper model partitioned the data set, fewer cases

were used to estimate each successive attribute's value. Thus, a stopping rule was needed to determine when there were not enough cases remaining in the data set to reliably estimate an attribute's likelihood ratio value. Our stopping rule required that there be at least ten cases remaining in the training sample before a question was evaluated.

Prior to the study a consultant from the ANN vendor (California Scientific Software) reviewed the APACHE II data base and recommended the ANN configuration used in this study. Attributes without missing data values were implemented as input neurons in numeric format and attributes with missing data values were implemented as input neurons in binary encoded form. This method for implementing attributes with missing values significantly increased the number of input and hidden neurons required by the ANN models. Since the output category was already encoded in binary form, it was implemented as a single neuron. The number of hidden neurons was calculated as the average of the number of input and output neurons. This method is customarily used to determine the initial number of hidden layer neurons and was recommended by the consultant. One hazard of binary encoding is that an attribute value may occur in the testing sample which does not also occur in the training sample. When this situation arose, attribute values in the testing sample were changed to the next lowest value which appeared in the training sample. Points to stop ANN training were assessed using the minimal root mean square and average error statistics.

The area under the ROC curve is an accepted statistic for measuring the accuracy of a dichotomous outcome variable and was used as the performance measure in this study.[15] The ROC curve area measured the ability of our models to discriminate between patients who lived and died. Thus, an ROC area of 1.00 denotes perfect discrimination and an ROC area of 0.500 denotes a lack of discriminatory ability. ROC curve areas in this study were calculated using the methods recommended by Hanley and McNeil[16, 17] ROC areas for different models were compared using the z statistic.

RESULTS

This study's results are presented in Tables 1 and 2. When missing data values were imputed as normal (Table 1), there were no statistically significant differences in ROC areas across all six models for two of the situations tested. These were : 100 training cases with 3 attributes and 600 training cases with 3 attributes. However, there were statistically significant differences in the other two situations. With 100 training cases and 11 attributes, both Bayesian models outperformed the logistic regression and ANN models. And, with 600 training cases and 11 attributes, all other models outperformed the ANN with 1 hidden layer.

In Table 1, the logistic regression models did not achieve maximum accuracy (ROC = 0.733) until they trained with 600 cases and 11 attributes. Their ROC areas in this situation were significantly greater than the other three situations tested. In contrast, the Bayes Simple model achieved its maximum accuracy (ROC = 0.749) with only 100 cases and 11 attributes. The maximum accuracy of the Bayes Proper model with 600 cases and 11 attributes (ROC = 0.732) was only slightly larger than with 100 cases and 11 attributes (ROC = 0.729). For both Bayes models, their accuracy with 100 or 600 cases and 11 attributes was significantly greater than in either of the other two situations with only 3 attributes. Both ANN models reached maximum accuracy with 600 cases and 11 attributes (ROC = 0.746 for 0 hidden layers and 0.692 for 1 hidden layer). And, the ANNs with 600 cases and 11 attributes were significantly more accurate than in the other three situations.

Table 1: No Missing Data (Imputed As Normal)

	100 Cases		600 Cases	
Model	**3 At**	**11 At**	**3 At**	**11 At**
Logist: Simple	0.663	0.645	0.627	0.733
Logist: Interact	0.663	0.663	0.627	0.733
Bayes: Simple	0.656	0.749	0.644	0.735
Bayes: Proper	0.654	0.729	0.644	0.732
ANN: 0 Hidn	0.663	0.662	0.650	0.746
ANN: 1 Hidn	0.631	0.600	0.650	0.692

Note: Cell values are ROC areas.

Table 2: Missing Data Allowed (Not Imputed)

	100 Case		600 Case	
Model	**3 At**	**11 At**	**3 At**	**11 At**
Logist: Simple	0.500	0.500	0.654	0.689
Logist: Interact	0.500	0.500	0.654	0.689
Bayes: Simple	0.653	0.730	0.657	0.730
Bayes: Proper	0.654	0.715	0.657	0.727
ANN: 0 Hidn	0.619	0.568	0.650	0.594
ANN: 1 Hidn	0.549	0.598	0.557	0.534

Note: Cell values are ROC areas.

When missing data values were not imputed (Table 2), some of the ROC areas were lower, but the basic

relationships between modeling techniques remained the same. With 100 training cases and 3 attributes, both Bayesian models and the ANN without hidden layers outperformed the other models. And, with 100 training cases and 11 attributes, both Bayesian models outperformed all other models. When there were 600 training cases and 3 attributes, all other models outperformed the ANN with 1 hidden layer. And, with 600 training cases and 11 attributes, both logistic regression and both Bayesian models outperformed both ANN models.

In Table 2, the logistic regression models did not achieve maximum accuracy (ROC = 0.689) until they trained with 600 cases and 11 attributes. Their ROC areas in this situation were significantly greater than the other three situations tested. The Bayes Simple model achieved its maximum accuracy (ROC = 0.730) with 11 attributes and either 100 or 600 cases. And, the Bayes Proper model reached maximum accuracy with 600 cases and 11 attributes (ROC = 0.727). With 600 cases and 11 attributes, the Bayes Proper model was more accurate than in situations with 100 or 600 training cases and 3 attributes; but it was no more accurate than when it trained with only 100 cases and 11 attributes. Both ANN models experienced problems when missing data values were allowed. The ANN with no hidden layers reached maximum accuracy with 600 cases and 3 attributes (ROC = 0.650) and the ANN with one hidden layer achieved maximum accuracy with 100 cases and 11 attributes. However, there were no situations for either ANN model in which they were significantly more accurate.

DISCUSSION AND CONCLUSIONS

This research design assumed that the major differences in model accuracy when missing data values were imputed (Table 1) would be determined by whether the models included interactions or not. However, the only significant differences (statistical or numeric) between models with and without interactions using the same technique occurred for ANNs and, in these, the models without hidden layers (no intreactions) exceeded the performance of models with hidden layers (with interactions). Thus, this research design assumption was not supported.

This design also assumed that the major differences in model accuracy when missing data values were allowed (Table 2) would be determined by whether the modeling technique was able to effectively manage missing data. The Bayesian models proved to be more accurate than the other techniques in these situations because they were able to more effectively manage missing data values. Thus, this research design assumption was supported.

When missing data values were imputed, the Bayesian models with 100 training cases responded to an increase in attributes (from 3 to 11) and did not require a corresponding increase in the number of training cases (from 100 to 600) to achieve maximum accuracy. In contrast, both the logistic regression and ANN models required larger training set sizes to increase their accuracy. Additionally, the performance of both the logistic regression and the ANN models suffered when missing data values were allowed while the Bayesian models had no significant difference in performance with or without missing values.

This study was a first test and we clearly need to develop better means for pre-configuring learning systems as well as better means for imputing or managing missing data values in ANNs. None of our models were able to achieve the 0.800 to 0.900 ROC areas which are commonly reported for classification models in the medical literature. Nonetheless, the Bayes Simple model did equal or exceed the accuracy of all other models in all situations tested. It achieved maximum accuracy with a smaller training sample, effectively managed missing data values, and did not require extensive set-up and configuration as was the case for the logistic regression and ANN models. For these reasons it deserves further attention and should be seen as a viable alternative to these other modeling techniques.

References

1. Durkin, J. Expert Systems: Design and Development. 1994 New York, NY: Macmillan Publishing Company.

2. Baxt, WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Annals of Internal Medicine 1991;115:843-848.

3. Pozen MW, D'Agostino RB, Mitchell JB, Rosenfeld DM, Guglielmino JT, Schwartz MI, Teebagy N, Valentine JM, Hood WB Jr. The usefulness of a predictive instrument to reduce inappropriate admission to the coronary care unit. Annals of Internal Medicine 1980;92(2 Pt 1):238-242.

4. Lette J, Colletti BW, Cerino M, McNamara D, Eybalin M-C, Levasseur A, Nattel S. Artificial intelligence versus logistic regression statistical modeling to predict cardiac complications after noncardiac surgery. Clinical Cardiology 1994;17:609-14.

5. Alemi F, Bhatt P, Eisenstein E, Fadlalla A, Stephens R, Butts J. Torturing data until they confess: A self learning bayesian expert system (B.E.St.). (Working Paper of the Collaborative Care Project). 1992 Cleveland, Ohio.

6. Chen S. Estimating Patient Severity Using A Bayesian Pattern Recognition Approach and A Neural Network Approach. Dissertation Abstracts International. 1993 (University Microfilm International).

7. Detrano R, Leatherman J, Salcedo EE, Yiannikas J, Williams G. Bayesian analysis versus discriminant function analysis: Their relative utility in the diagnosis of coronary disease. Circulation 1986; 73:970-977.

8. Morise AP, Dival RD, Detrano R, Bobbio M, Diamond, GA. Comparison of logistic regression and bayesian-based algorithms to estimate posttest probability in patients with suspected coronary artery disease undergoing exercise ECG. Journal of Electrocardiology 1992;2589-99.

9. Chiang CJ, Bernstein AD, Parsonnet V. A probability-based expert system for diagnosing pacemaker-related complications. Computers In Cardiology 1994. (pp. 89-92): IEEE Computer Society Press. Los Alimitos, CA.

10. Chard T. Self-learning for a bayesian knowledge base: How long does it take for the machine to educate itself? Methods of Information in Medicine 1987;26:185-88.

11. Alemi F, Rice J, Hankins R. Predicting in-hospital survival of myocardial infarction. Medical Care 1990;28:762-775.

12. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: A severity of disease classification system. Critical Care Medicine 1985;13:818.

13. Shavlik JW, Mooney RJ, Towell GG. Symbolic and neural network algorithms: An experimental comparison. Machine Learning 1991;6:111-43.

14. Baxt WG. A neural network trained to identify the presence of myocardial infarction bases some decisions on clinical associations that differ from accepted clinical teaching. Medical Decision Making 1994;July-Sept:217-22.

15. Iezzoni LE. Risk Adjustment for Measuring Health Care Outcomes. 1994 Ann Arbor, MI: Health Administration Press.

16. Hanley J, McNeil B. The meaning and use of the areas under a receiver operating curve. Diagnostic Radiology 1982;143:29-36.

17. Hanley JA and McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same data. Radiology 1983;148:839-43.