Selecting Modeling Techniques for Outcome Prediction: Comparison of Artificial Neural Networks, Classification and Regression Trees, and Linear Regression Analysis for Predicting Medical Rehabilitation Outcomes

Deborah K. W. Walters[1], Ph.D., Richard T. Linn[2], Ph.D., Margaret Kulas[2], B.S., Elisabeth Cuddihy[1], M.S, Chonghua Wu[1], M.S., Carl V. Granger[2], M.D.
[1]Computer Science and Engineering and [2] Rehabilitation Medicine, SUNY at Buffalo, Buffalo, N.Y.

Abstract

A multitude of techniques exists for modeling medical outcomes. One problem for the researcher is how to select an appropriate modeling technique for a given task. This paper addresses the problem through: an analysis of the strengths and weaknesses of three techniques; and, a case study in which the three techniques are applied to the task of predicting medical rehabilitation outcomes. The three techniques selected where linear regression analysis (LRA), classification and regression trees (CART) and artificial neural networks (ANN). The analysis illustrates that when the relationship between the independent and dependent variables is a linear one, that LRA is adequate. However, when a nonlinear relationship exists, CART or ANN analysis will yield better models. The results of the case study show that the ANN model is more accurate than both LRA and CART in predicting the discharge motor FIM from admission data for stroke patients admitted to medical rehabilitation facilities. However, the increased accuracy comes at an increase in the computational cost of the model, thus a decision about which technique to use must be made by weighing the increased accuracy against the increased cost.

Introduction

The selection of a modeling technique to use for a given medical task is a nontrivial problem as there are many techniques from which to choose. In particular, when the goal is to predict medical outcomes, the literature shows that many techniques have been used. One way to approach the problem is to look at previous research, especially those that compare the performance of different modeling techniques [1]. This paper presents an analysis of three different techniques that is designed to aid a researcher in determining which of the three techniques would be most appropriate for a given problem. In addition, a case study is presented showing the use of the three techniques in predicting the discharge motor FIM score of stroke patients who were admitted to medical rehabilitation facilities.

The three techniques selected were linear regression analysis (LRA), classification and regression trees (CART) and artificial neural networks (ANN). LRA was selected as it is a simple, readily available and commonly understood technique. The use of any more complex technique should be justified on the grounds of improved performance in comparison to LRA. CART was selected because of it's demonstrated success in modeling function related groups (FRG's) in medical rehabilitation [2, 3], and because of it's relative potential for other types of medical outcome modeling. ANN's were selected because of the growing interest in their use in modeling medical systems [1,4,5,6].

Analysis of Techniques

There are many reasons why a model may have low predictive value. It could be that data incorporated into the model is insufficient to predict the outcome, in which case no model will be able to correctly predict the outcome. Or it could be that the modeling technique used is not the most appropriate for the task and that accuracy can be increased through the use of a more appropriate model. One way in which a technique may be inappropriate is that the problem may be nonlinear and the technique may be linear.

To illustrate this, a 2-category 2-D classification problem can be used. Consider a simple problem where the goal is to predict for a given individual, which of two possible outcomes will occur. Without loss of generality, further assume that the outcome is a function of only two independent variables, X and Y. Then the goal is to develop a model where the outcome can be predicted from the values of X and Y for a given individual.

A model for this prediction will consist of dividing the X-Y space into two regions: one for which most individuals have the first outcome, and the other for which most have the second outcome. When the two regions can be adequately defined using a single straight line, then the problem is a linear one, and can be adequately modeled by a linear technique such as LRA. In general, the line is referred to as the "decision surface". If the problem is not linear, then a nonlinear technique such as CART or ANN must be used. For the 2-category 2-d classification problem, CART divides the XY space into rectangular regions, and then groups the rectangles into the two regions, and the step-like boundary between the regions is the decision surface. By creating smaller and smaller rectangles, the CART decision surface can become as close as desired to any arbitrary actual decision surface.

An ANN can create a decision surface with an arbitrary shape. Consider, for example, a situation where the shape of the optimal decision surface is a circle. The ANN decision surface can be a circle, while the CART decision surface will approximate the circle by creating a multitude of small rectangles. Consider another example, where the optimal decision surface is a diagonal straight line. The LRA and the ANN decision surfaces will be the diagonal straight line, but the CART decision surface will again be approximated by the boundary between many small rectangles. Thus the ANN decision surface has the potential to be simpler than the CART decision surface.

If the problem is linear, then LRA will be adequate, while if the problem is nonlinear, CART or ANN would be better choices, as the LRA model will be limited in it's accuracy. How would one know if a given problem were nonlinear? Could one simply create a NLA model to see if it was accurate? The answer is "no" because low accuracy in a linear model could be due to a non-linearity, to noise or to a combination of the two. It is true that for a given problem, if a LRA model has low accuracy and an ANN or CART model gives high accuracy, it could be inferred that the problem has a non-linear component. But low accuracy alone from an LRA model is not enough to make the inference of non-linearity in the problem.

Why not then just always use ANN models? As discussed below, creating an ANN model requires making lots of decisions about the types and values of parameters to use. Poor choices of these values can lead to poor performance of the model. And searching for the optimal values of the parameters can be computationally expensive. Thus the creation of good ANN models requires significant knowledge and can be expensive.

Methods

Models were created using the three selected techniques and their performance and costs were compared. The problem modeled was the prediction of the discharge motor FIM scores of stroke patients admitted to medical rehabilitation facilities. Data records were obtained from the Uniform Data System for Medical Rehabilitation (UDSmr) for over 65,000 first admission, stroke patients who were discharged from rehabilitation facilities in 1995. A random sample of 20,000 records was formed form the entire data set. Half of these records were used as the training set for the models. The other half of the records was used as the test set, and the reported accuracies are those for the models applied to the test set.

The data records contained the following information: age, gender, onset time, admission motor FIM scores, admission cognitive FIM scores, marital status, ethnicity, pre-admission living arrangements, primary payment source, region of facility and discharge motor FIM scores. The motor FIM scores were reported for each of thirteen categories such as eating and walking. The cognitive FIM scores were reported for each of six categories including items such as problem solving and memory.

One of the motivations for this research was to develop predictive models that would be of use to clinicians, patients and their families. When a patient enters a rehabilitation facility, it would be useful to be able to predict what the motor functioning of the patient would be when they are discharged. Thus it would be useful to be able to predict the discharge motor FIM score. There are several different forms that such a prediction could take. One form would be an exact prediction where the model was judged to be accurate only if the model prediction was identical to the actual discharge FIM score. However, given the variability in this type of problem, it is unlikely that very high accuracies could be achieved for an exact prediction. A more reasonable form for a prediction would require that the predicted value fall within a given range of the actual value. The choice of the given range is important, as any level of accuracy is possible as long as the range is large enough. For these experiments five types of predictions were selected based on their potential usefulness to the clinician. The ranges differ in their extent, and thus the accuracies of a given model will vary for the different predictions.

The five predictions utilized were:

a) Maximal Accuracy Prediction: (+/-5 points) Predicted value must lie within a narrow range of the actual value.
b) Minimal Level Prediction: (-5 or greater) Actual value is no less than 5 points below predicted value.
c) Significant Assistance Prediction: (52 or greater) Predicts whether patient will have score of 52 or greater. A score of 52 could be achieved by having a score of 4 (indicating significant assistance required) in each of the 13 categories.
d) Moderate Assistance Prediction: (65 or greater) Predicts whether patient will have score of 65 or greater. A score of 65 could be achieved by having a score of 5 (indicating moderate assistance required) in each of the 13 categories.
e) Little Assistance Prediction: (78 or greater) Predicts whether patient will have score of 78 or greater. A score of 78 could be achieved by having a score of 6 (indicating little assistance required) in each of the 13 categories.

For each type of prediction, models were created using the three techniques. In addition a simple model was created for each of the prediction types to give a base level performance that the other models should be expected to surpass.

The LRA was performed using SPSS. The CART analysis was performed using an implementation of the algorithm defined in [7]. The ANN models were generated using two different software packages: the ANN I models used the ASPIRIN package [8] and the ANN II models used MatLab.

The Simple I model was created by determining the average gain in motor FIM for patients in the test set, and then simply adding that value to the admission motor FIM to get a predicted value. The Simple II models were created by determining if the majority of patients in the test set had discharge scores less than the cut-off value in the prediction or not, and then predicting that each patient falls within the majority group.

Two CART models were created: CART I predicted the total discharge motor FIM score; and CART II used a separate CART model to predict each of the 13 categories of the motor FIM, and then summed the predictions to get a final prediction.

Results

Initially all models were developed using all of the independent variables. Preliminary results showed that only three of the independent variables were contributing significantly to the models: admission motor FIM, admission cognitive FIM and age. In LRA and CART, this result was apparent due to the low correlation coefficients for some independent variables. In the ANN models this result was seen through low weights between the input layer and the next layer for some independent variables.

All of the models reported here used as their independent variables just admission motor FIM, admission cognitive FIM and age.

Maximal Accuracy Prediction Results

The following table shows the accuracy for six models for the Maximal Accuracy Prediction.

Model	**Accuracy**
Simple I	35.7%
Linear Regression	37.8%
CART I	40.6%
CART II	42.2%
ANN I	42.9%
ANN II	43.6%

Minimal Level Prediction Results

The following table shows the accuracy for five models for the Minimal Value Prediction.

Model	**Accuracy**
Linear Regression	68.8%
CART I	70.0%
CART II	65.8%
ANN I	73.2%
ANN II	74.0%

The accuracies for each model for the Minimal Value Prediction are higher than those for the same model for the Maximal Accuracy Prediction. This result is to be expected since the range of accepted predictions is much larger for the Minimal Value Prediction.

Significant Assistance Prediction Results

The following table shows the accuracy for five models for the Significant Assistance Prediction.

Model	Accuracy
Simple II	72.9%
Linear Regression	85.2%
CART I	85.3%
ANN I	85.6%
ANN II	85.5%

Moderate Assistance Prediction Results

The following table shows the accuracy for five models for the Moderate Assistance Prediction.

Model	Accuracy
Simple II	51.1%
Linear Regression	80.2%
CART I	80.6%
ANN I	81.0%
ANN II	81.3%

Little Assistance Prediction Results

The following table shows the accuracy for five models for the Little Assistance Prediction.

Model	Accuracy
Simple II	78.3%
Linear Regression	82.6%
CART I	84.6%
ANN I	82.2%
ANN II	84.9%

For all of the predictions, the Simple Models had the lowest accuracies. The CART I models had accuracies higher than the LRA models for all predictions. The ANN II models all had accuracies higher than the CART models for all predictions.

The following table shows the percentage increase in accuracy of the best CART model and the best ANN model versus the LRA model for each prediction. The last column shows the added percent increase in accuracy that the best ANN model had over the best CART model. Thus for Maximal Accuracy Prediction, an 11.6% increase in accuracy occurs with the use of the best CART model as compared to the LRA model, and further 4.7% increase occurs with the use of the best ANN model.

Prediction	CART vs. LRA	ANN vs. LRA	ANN vs. CART
Maximal Accuracy	11.6%	15.3%	4.7%
Minimal Level	01.7%	07.6%	5.9%
Significant Assistance	00.4%	00.5%	0.1%
Moderate Assistance	00.5%	01.4%	0.9%
Low Assistance	02.4%	02.8%	0.4%

Discussion

Two different ANN experts using two different software packages created the ANN models. The different results of the two models illustrates the fact that it is generally not possible to find a single “best” ANN model for a given problem. Another expert would probably produce yet other models.

The relatively small improvements in the accuracies of the CART and the ANN models versus those of the LRA models suggests that there is not a large non-linear component to this problem.

The ANN II model has the best accuracy of all the models. Does this suggest that artificial neural network models should be used for all medical outcome predictions? Does this even suggest that the ANN II models should be used to develop a commercial product for predicting discharge motor FIM scores? The answer to both questions is, “Not necessarily.” The issue is that unlike LRA modeling where there is only a single correct linear regression model for a given data set, there is no single correct neural network model. In creating a neural network model there are a large number of decisions that must be made, including:

1. Which learning algorithm should be used?
2. Which architecture?
3. How many layers?
4. Which activation functions?
5. Which input representations?
6. Which output representations?
7. What learning rate?
8. How long to train?

The large number of decisions means that there is a very large space of possible neural networks for a given data set. In creating a neural network model the goal is to find the best network by searching through this large space. To search the space requires that one implements multiple networks and then tests their performance to find the best network. Since training each neural network can require significant computational resources, there can be a significant computational cost to searching the space. In addition there can be significant costs in terms of personnel time to perform such a search. While there are some software packages that assist the implementers in searching the space, no package exists which enables a search of the entire space. Experienced neural network researchers have developed several rules of thumb that can aid the search. However, even an expert cannot determine the correct choices for a given problem since the best choice depends upon the structure of the data, which is not known in advance. In summary, creating a good artificial neural network can be costly.

While the CART analysis requires a few decisions made on the part of the implementers, the space of possible models for a given data set is much smaller. In addition, the creation of CART models requires fewer computational resources than ANN models. Thus the cost of implementing a CART model is much less than the cost of implementing an ANN model.

The LRA model is the least expensive of all models to implement.

In choosing a modeling technique, one must weigh the costs of the techniques against the accuracies of the techniques. While it may be cost effective for the Minimal Level prediction to use an ANN model to gain an additional 5.9% in accuracy, it may not be cost effective to use an ANN or a CART model for the Maximal Assistance Prediction.

References

1. Lang EW, Pitts LH, Damron SL, Rutledge R. Outcome after severe head injury: Analysis of prediction based upon comparison of neural network versus logistic regression analysis. *Neurological Research* 1997; **19:** 274-280
2. Stineman MG, Goin JE, Tassoni CJ, Granger CV, Williams SV. Classifying rehabilitation inpatients by expected functional gain. *Medical Care* 1997; **35:** 963-973
3. Stineman MG, Hamilton BB Goin JE Granger CV Fiedler RC. Functional gain and length of stay for major rehabilitation impairment categories. *American Journal of Physical Medicine and Rehabilitation* 1996; **75:** 68-78
4. Grisby J, Kooken R, Hershberger J. Simulated neural networks to predict outcomes, costs, and length of stay among orthopedic rehabilitation patients. *Arch. Phys. Med. Rehabil.* 1994; **75:** 1077-1082
5. Oczkowski WJ, Barecca S. Neural network modeling accurately predicts functional outcome of stroke survivors with moderate disabilities. *Arch Phys Med Rehabil* 1997; **78:** 340-345
6. Sonoda S, Chino N, Domen K Saitoh E. Changes in impairment and disability from the third to the sixth month after stroke and its relationship evaluated by an artificial neural network. *American Journal of Physical Medicine and Rehabilitation* 1997; **76:** 395-400
7. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression tress. Belmont, CA: Wadsworth, Inc., 1984
8. ASPIRIN, The Rand Corporation.