MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press
© 1998 IMIA. All rights reserved

Using Classification Tree and Logistic Regression Methods to Diagnose Myocardial Infarction

Christine L. Tsien[a,b], Hamish S. F. Fraser[a,c], William J. Long[a], R. Lee Kennedy[d]

[a]*Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA*
[b]*Division of Health Sciences and Technology, Harvard Medical School, Boston, MA, USA*
[c]*Division of Clinical Decision Making, Tufts-New England Medical Center, Boston, MA, USA*
[d]*Department of Medicine, Sunderland District General Hospital, Sunderland, UK*

Abstract

Early and accurate diagnosis of myocardial infarction (MI) in patients who present to the Emergency Room (ER) complaining of chest pain is an important problem in emergency medicine. A number of decision aids have been developed to assist with this problem but have not achieved general use. Machine learning techniques, including classification tree and logistic regression (LR) methods, have the potential to create simple but accurate decision aids. Both a classification tree (FT Tree) and an LR model (FT LR) have been developed to predict the probability that a patient with chest pain is having an MI based solely upon data available at time of presentation to the ER. Training data came from a data set collected in Edinburgh, Scotland. Each model was then tested on a separate Edinburgh data set, as well as on a data set from a different hospital in Sheffield, England. Previously published models, the Goldman classification tree[1] and Kennedy LR equation[2], were evaluated on the same test data sets. On the Edinburgh test set, results showed that the FT Tree, FT LR, and Kennedy LR performed equally well, with ROC curve areas of 94.04%, 94.28%, and 94.30%, respectively, while the Goldman Tree's performance was significantly poorer, with an area of 84.03%. The difference in ROC areas between the first three models and the Goldman model is significant beyond the 0.0001 level. On the Sheffield test set, results showed that the FT Tree, FT LR, and Kennedy LR ROC areas were not significantly different ($p >= 0.17$), while the FT Tree again outperformed the Goldman Tree ($p = 0.006$). Unlike previous work[3], this study indicates that classification trees, which have certain advantages over LR models, may perform as well as LR models in the diagnosis of patients with MI.

Keywords

Classification Trees; Decision Trees; Logistic Regression; Myocardial Infarction; Machine Learning

Introduction

A major part of emergency medicine is accurately diagnosing or ruling out myocardial infarction (MI) in patients who present to the Emergency Room (ER) complaining of chest pain. Early diagnosis not only enables better medical management of patients, but also prevents unnecessary stress and anxiety. Furthermore, the ability to rule out MI in a patient with chest pain translates into cost savings. Patients who are safe to go home do not need to be unnecessarily admitted to the hospital, and those who have less severe medical problems can be admitted to a general ward instead of to the cardiac care unit (CCU).

One approach to early diagnosis of MI is to use machine learning techniques to develop appropriate models. The goal of machine learning could be to devise either a simple flow chart type of clinical decision aid, or a more compli-cated equation that determines the probability of a given patient having MI. A decision aid for this task may be a paper flow chart or a calculator. Alternatively, classifica-tion trees or mathematical models may be implemented as computerized calculators, for example, in a web page.

In this study, a flow chart type of model was developed using classification tree techniques. An equation type of model was also developed using logistic regression (LR) methods. For each of these techniques, the information used to determine a patient's condition includes clinical and electrocardiographic (ECG) data available at the time the patient presents in the ER.

Previous work in decision aids for MI diagnosis[1-6] have been developed but have not achieved general use. Reasons cited include difficulty in generalizability of results to other hospitals, and the impractical nature of some tree models.

This study explores development of a flow chart type of aid for early diagnosis of MI that performs at least as well as LR models, which have previously been thought to perform better. Following presentation of the methods used, the models built will be described and their performance compared with that of previous models. Finally, some of the more interesting issues will be discussed.

Methods

Chest Pain Data

The chest pain data was originally collected by Kennedy *et al.*[2] to look at the derivation and evaluation of LR models. For the current project, 1752 data cases were used, corresponding to 1252 patients presenting with chest pain to the Edinburgh Royal Infirmary in Scotland, and 500 patients presenting with chest pain to the Northern General Hospital in Sheffield, England.

For each data case, there are 45 attributes available. Table 1 lists these attributes.

Both classification tree and LR model building used a subset of the available Edinburgh data (630 cases; 23% occurrence of MI), while the remainder (622 cases; 21% MI) was reserved for model evaluation. The entire Sheffield data set (500 cases; 31% MI) was reserved for evaluating the models on cases from a different hospital and region.

Table 1 - Patient attributes collected

age	smoker	ex-smoker
family history of MI	diabetes	high blood pressure
lipids	retrosternal pain	chest pain major symptom
left chest pain	right chest pain	back pain
left arm pain	right arm pain	pain affected by breathing
postural pain	chest wall tenderness	sharp pain
tight pain	sweating	shortness of breath
nausea	vomiting	syncope
episodic pain	worsening of pain	duration of pain
previous angina	previous MI	pain worse than prev. Angina
crackles	added heart sounds	hypoperfusion
heart rhythm	left vent. hypertrophy	left bundle branch block
ST elevation	new Q waves	right bundle branch block
ST depression	T wave changes	ST or T waves abnormal
old ischemia	old MI	sex

Tree Building and Comparison

Classification trees can be generated by machine learning algorithms used to classify new data using a tree structure derived from a sample of "training" data of known classification. Each "datum" consists of several attributes (e.g., chest pain attributes), and one class label (e.g., MI or non-MI). The trees are built by looking for regularities in the data with which to separate the data by class.

The classification tree built for this project used Quinlan's C4.5 program[7] written for Unix systems. C4.5 takes as input a file specifying the available attributes and their value type, as well as the possible classifications for each sample. Training data cases are provided in a separate file, and optional test items are provided in a third file.

Several options are available for modifying tree-building behavior. The options experimented with include: *m*, the minimum number of cases needed in at least two outcomes of a tree node in order to include that node while creating the tree; c, the confidence level of the predicted error rate on each leaf and each subtree, used to find the upper limit on the probability of error at a leaf or subtree during pruning; and t, the number of trees to be grown by partitioning the given training set, the best of which is then selected.

Clinical judgment was used to select the final tree (called FT Tree) of those with the best numerical results on training data. This involves determining whether proposed attributes are themselves consistent with the goal; whether attributes within a given branch make sense with respect to each other; and in cases where an attribute is repeated more than once in the same branch, whether there is clinical plausibility for this. The FT and Goldman[1] Trees were then implemented as C programs to facilitate comparisons on the same test sets. The Long Tree[3] was not compared in this manner because several of its attributes were not available in the Edinburgh and Sheffield data.

LR Model Building and Comparison

LR is a non-linear classification method which uses a set of samples of known classification to derive coefficients for an equation that calculates the probability that a new case is of a certain class. This equation is written as:

$$\text{Probability} = 1/(1+\exp[-(\beta_0 + \Sigma\beta_i X_i)]) \quad (1)$$

where β_0 is a constant term, β_i terms are the derived coefficients, and X_i terms are the values of the attributes used to determine the cases classification (0 or 1 for dichotomous type; integers, for example, for continuous type).

In general, the time-consuming and difficult aspect of building an LR model is deciding which attributes to include in the model and which to exclude. Recalling that there are 45 available attributes related to chest pain in this particular data set, choosing a small but meaningful subset could pose a daunting task. However, the approach taken in this study has been to allow C4.5 to select the optimum variables. This simplified the job of building an LR model using the JMP statistical package (SAS Institute, Carey, NC). Age and duration were used as continuous variables, although C4.5 provides useful methods to dichotomize variables if required. The model developed here (FT LR) and the Kennedy LR model [2] were then implemented as C programs to facilitate comparisons on the same test sets.

Calculations used for Comparison

The calculations used for comparing the performance of classification tree and LR models include sensitivity, specificity, positive predictive value (PPV), accuracy, and area under the receiver operating characteristic (ROC) curve. Sensitivity is the number of correct model-diagnosed patients with MI, divided by the number of gold-standard patients with MI. Specificity is the number of correct model-diagnosed patients without MI, divided by the number of gold-standard patients without MI. PPV is the number of correct model-diagnosed patients with MI, divided by the number of all patients model-diagnosed with MI (correctly or incorrectly). Accuracy is the number of patients with a correct model-diagnosis divided by the total number of patients. For classification trees, ROC curves were derived by first assigning to each tree leaf the probability of having MI for a patient whose case arrives there. These probabilities are based upon the ratio of MI to non-MI patients in the training data that fall into each leaf. The threshold for considering a case to be MI or non-MI was then set at each leaf probability value. The resulting sensitivity-specificity pairs were plotted on a grid of sensitivity versus (1-specificity) to obtain the ROC curve. For each LR model, twelve threshold values (0, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100%) were used for determining whether to diagnose a case as MI or non-MI; the sensitivity-specificity pairs were similarly plotted. The area and standard error (SE) under each ROC curve was calculated by the Hanley-McNeil method[8].

Results

Classification Trees and Comparisons

Figure 1 shows the FT Tree. For each leaf, the number of correctly classified cases out of the total number of cases in that leaf are listed. (Fractions arise due to tree pruning in model building.) The values for the variables of tree-building used were: $m = 5$ minimum number of cases in at least two branches of a node, $c = 15\%$ confidence level, $t = 10$ trees built using partitioning of the given training set. Table 2 presents the results of the FT, Goldman, and Long Trees, each on their *own* test sets. Table 3 compares the attributes of the FT, Goldman, and Long Trees (those above the line are similar among the trees). Approximately half of the Goldman Tree and one third of the Long Tree attributes are similar to those in the FT Tree.

```
ST elevation = 1: 1 (40.7/49.0 = 83.1%)
ST elevation = 0:
| New Q waves = 1: 1 (4.1/7.0 = 58.6%)
| New Q waves = 0:
| | ST depression = 0: 0 (329.4/345.0 = 95.5%)
| | ST depression = 1:
| | | Old ischemia = 1: 0 (3.2/6.0 = 53.3%)
| | | Old ischemia = 0:
| | | | Family history of MI = 1: 1 (6.8/11.0 = 61.8%)
| | | | Family history of MI = 0:
| | | | | age <= 61 : 1 (4.0/8.0 = 50.0%)
| | | | | age > 61 :
| | | | | | Duration of pain (hours) <= 2 : 0 (14.1/22.0 = 64.1%)
| | | | | | Duration of pain (hours) > 2 :
| | | | | | | T wave changes = 1: 1 (7.0/10.0 = 70.0%)
| | | | | | | T wave changes = 0:
| | | | | | | | Right arm pain = 1: 0 (3.4/5.0 = 68.0%)
| | | | | | | | Right arm pain = 0:
| | | | | | | | | Crackles = 0: 0 (3.0/8.0 = 37.5%)
| | | | | | | | | Crackles = 1: 1 (4.9/9.0 = 54.4%)
```

Figure 8 - FT Tree, the final tree selected

Table 2 - Goldman, FT, and Long Trees on OWN test sets

	Goldman	FT Tree	Long
Sensitivity =	90.9%	81.4%	66.1%
Specificity =	69.7%	92.1%	85.8%
PPV =	35.4%	72.9%	68.3%
Accuracy =	73.1%	89.9%	80.1%

The area under the ROC curve for the FT Tree on its test set was 94.04% (SE = 0.72%), while that of the Long Tree on its own test set was 86%[3]. Since Goldman did not report an area, it was calculated by running the Goldman Tree on the Edinburgh test set. The resulting area was 84.03% (SE = 2.28%). Figure 2 shows the ROC curves for FT and Goldman run on Edinburgh data; they have significantly different ROC curve areas ($p < 0.0001$).

Since the FT Tree had been trained on Edinburgh data, a more rigorous comparison is to run the trees on previously unseen data from a different hospital (Sheffield). Figure 3 shows the resulting ROC curves. The difference in areas for the FT (89.61%, SE = 1.07%) and Goldman (83.85%, SE = 2.04%) Trees was again significant ($p = 0.006$).

Table 3 - Attributes of Goldman, FT, and Long Trees

Goldman	FT	Long
ST elevation or Q waves	ST elevation New Q waves	ST change Q waves
Duration	Duration	
ST or T wave	T wave	T wave
Shoulder, neck, arm	Right arm	Arm,neck,shoulder
Age	**Age**	**Age**
Local Pressure	ST depression	Stomach pain
Previous angina	Old ischemia	Sex
Left shoulder	Family history	Systolic BP
Pain worse	Crackles	Heart rate
Diaphoresis		Rapid/skipping beats
		Chest pain
		History of MI
		Nitroglycerin use
		Shortness of breath
		Fainted, dizzy, lightheaded

Logistic Regression Models and Comparisons

The FT LR model is presented in Table 4, along with the coefficients of the Kennedy LR[2] and Long LR[3] models for comparison. The Kennedy and FT equations use several of the same chest pain attributes, whereas the Long and FT models do not. Long reported an ROC area of 89%[3]. The Kennedy and FT LR models were each tested on Edinburgh and Sheffield data sets. The ROC area for Kennedy on the Edinburgh data was 94.30% (SE = 0.92%), while that for FT LR was 94.28% (SE = 1.16%). On Sheffield data, the ROC area for Kennedy was 91.25% (SE = 1.32%), while for FT LR was 89.28% (SE = 1.59%). No difference exists between the models ($p = 0.50$ and $p = 0.17$, respectively).

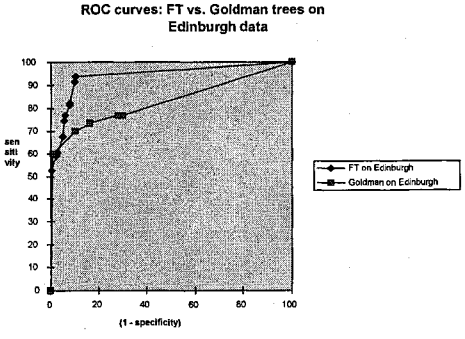

Figure 9 - Goldman Tree (area = 84.03%) vs. FT Tree (area = 94.04%) on Edinburgh data, $p < 0.0001$

Classification Trees versus Logistic Regression

The FT Tree performed similarly to the best LR model on Edinburgh data ($p = 0.41$) and Sheffield data ($p = 0.17$). The Goldman Tree performed appreciably lower than that of the other models on both data sets. Table 5 lists the ROC curve areas for each model by test set. Figure 4 shows the ROC curves for all models evaluated on Sheffield data.

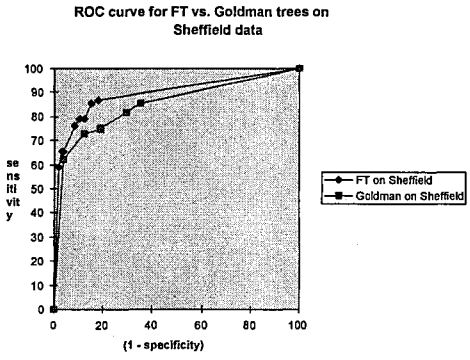

Figure 10 - Goldman Tree(area = 83.85%) vs. FT Tree (area = 89.61%) on Sheffield data, p = 0.006

Discussion

Table 4 - Comparison Kennedy, FT, and Long LR models

	Kennedy	FT LR	Long LR
Constant	-3.07	-2.14	
ST elevation	3.16	2.96	
New Q waves	1.37	2.00	
ST depression	1.95	1.76	
LV Failure (Crackles)	1.54	0.807	
Old ischemia		-0.86	
Family hx		0.43	
Age		-0.016	
Duration		-.0046	
T wave		0.805	
Right arm		-0.22	
Vomiting		0.68	
Hypoperfusion		0.47	
Chest pain #1 Sx			0.71
Chest pain/24h			1.00
T wave nl/flat			1.13
Nitro use			0.51
Previous MI		0.42	
STchange nl/flat			0.77
STchange normal			0.83

Table 5 - Comparisons of Areas under the ROC Curve

Model:	Edinburgh data	Sheffield data
FT Tree	94.04%	89.61%
Goldman Tree	84.03%	83.85%
FT LR	94.28%	89.28%
Kennedy LR	94.30%	91.25%

The results have indicated that the FT Tree can perform as well as LR models. This differs from an earlier report[3] which suggested that LR models for early diagnosis of MI perform better than classification tree models.

The FT Tree had better results than both Goldman and Long Trees. Additionally, the FT Tree is relatively small and clinically reasonable. The Goldman Tree is also relatively small, but performed less well. Moreover, a few of the paths through the Goldman Tree, especially those that ask about age or duration of pain multiple times before reaching a leaf, seem less clinically appropriate. The Long Tree is quite large, did not perform as well as the FT Tree, and also has several paths that seem clinically inappropriate (e.g., comparing pulse rate to the thresholds of 77 beats per minute and 89 beats per minute in consecutive nodes).

It would have been preferable to implement the Long[3] and Selker[4] models also. (It should be noted that the Selker model is intended to diagnose unstable angina as well as MI.) Effectively dealing with missing values is one possible approach to be taken in the future.

Previous work[2] indicated that the difficulty in deriving a good LR model is in choosing which attributes to include or exclude from the model. Given that the FT LR model performed as well as the Kennedy model, one use of classification trees may be to select attributes for LR models. Another use could be to decide breakpoint values for continuous variables should dichotomous values be required. Preliminary experimentation of building an LR model with the age and duration of pain attributes as dichotomous values (breakpoints at 61 years, and 2 hours, respectively), did not appear to have any significant change in the resulting ROC curve areas.

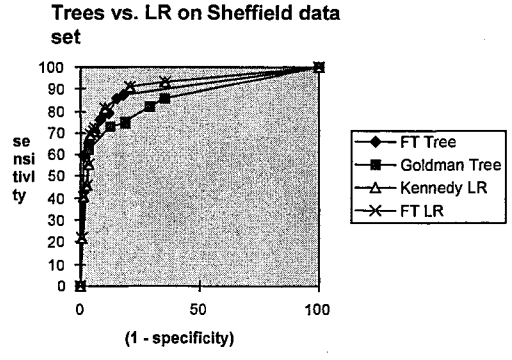

Figure 11 - Trees vs. LR for Sheffield data

For all ROC curve area calculations presented here, the Hanley-McNeil method was used. This method, however, may not be as accurate as the Dorfman and Alf maximum likelihood estimation program, or the slope and intercept of the original data when plotted on binormal graph paper. In calculating ROC areas for LR models, this may not present a problem since many threshold values can be used to derive many sensitivity-specificity pairs. It may present more of a problem for classification trees, however, in which the number of sensitivity-specificity pairs is limited by the number of leaves in the tree. With fewer points on the ROC curve, underestimation of the actual area and thus performance may be accentuated for classification trees.

Even with this underestimation, however, the FT Tree performs competitively. A sequential Bayes method[6] and neural networks[4, 5] have also been explored for MI diagnosis. The

Bayes method had promising results, but requires using a heuristic method to handle interdependence of data. While the neural network models have also had promising results, they remain a "black box" model type not easily understandable to the clinician and thus less likely to be accepted into general use. LR models are likewise less understandable to clinicians than flowchart-type decision trees. All of the methods share the difficult task of trying to produce models that are generalizable to other hospitals. Additional testing of existing models on various data sets may provide useful suggestions for future directions.

Conclusion

The increasing availability of hospital information systems provides an ideal environment for implementing decision tools. Models may also be included in web pages and ECG machines. The FT Tree, for example, can be implemented as a one-page flow chart that can be inserted into the patient chart. Prospective, randomized studies in a clinical setting are required to fully validate such decision aids.

Acknowledgments

The authors would like to thank Isaac Kohane, MD PhD and Peter Szolovits, PhD for advice. This work was supported in part by the National Library of Medicine and Howard Hughes Medical Institute.

References

[1] Goldman L, Weinberg M, Weisberg M, Olshen R, Cook E, Sargent R, Lamas G, Dennis C, Wilson C, Deckelbaum L, Fineberg H, Stiratelli R. A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. *N Engl J Med* 1982; (307): 588-596.

[2] Kennedy RL, Burton AM, Fraser HS, McStay LN, Harrison RF. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models. *EHJ* 1996; (17): 1181-1191.

[3] Long WJ, Griffith JL, Selker HP, and DAgostino RB. A comparison of logistic regression to decision-tree induction in a medical domain. *Comput Biomed Res* 1993; (26): 74-97.

[4] Selker HP, Griffith JL, Patil S, Long WJ, DAgostino RB. A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. *Journal of Investigative Medicine* 1995; (43): 468-476.

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

[6] Jonsbu J, Aase O, Rollag A, Liestol K, Erikssen J. Prospective evaluation of an EDB-based diagnostic program to be used in patients admitted to hospital with acute chest pain. *EHJ* 1993; 14:441-6.

[7] Quinlan JR. *C4.5 Programs for machine learning.* San Mateo: Morgan Kaufmann Publishers, 1993.

[8] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; (143): 29-36.

Address for correspondence

Christine L. Tsien
545 Technology Square
NE43-420
Cambridge
MA 02139
USA
chris@medg.lcs.mit.edu.