UNIVERSITY OF TEXAS AT SAN ANTONIO

Assessment of Data Mining Models with Beijing Transportation Data

Liang Jing

May 2008

1 ABSTRACT

A data set from an Internet questionnaire about transportation condition in Beijing is analyzed with the help of data mining techniques. The data set is preprocessed by cleaning the original data, examining and combining related variables. Then several variables are studied as the response variables in the purpose of finding other variables which have impact on them. A variety of models are considered, including linear models, generalized linear models, generalized additive models, tree, and neural networks. Stepwise variable selection procedure is conducted for selected models, and model efficiency is evaluated by using different scenarios as well as cross-validation. For data sets with small sample size, bootstrapping is used to enlarge the sample size and help identifying the potential power of the models. By carefully comparing the prediction performance of these models, several conclusions, regarding to model complexity, "overfitting", modeling assumptions and etc, are drawn.

2 INTRODUCTION

Transportation is a serious problem for every city in the world, especially for big cities like Beijing which has over 10 million population. The citizens who suffer this problem everyday are most experienced. Thus, collecting information from them is the best way to learn the problems and find solutions. Due to the large number of both variable size and sample size, data mining techniques are necessary and important for analyzing the collected data and producing useful results.

The data set used in this study is from an Internet questionnaire about transportation condition in Beijing, conducted in October 2004 by Social Psychological Institution of People's University of China. There are 26 questions and some of them consist of sub-questions which make totally 68 questions. These questions are related to personal background, driving history, driving habits, rating to transportation system, and etc (a complete list of the questions is attached).

Except for Question 3, 18, 20, 25 and 26, answers for other 63 questions are single choice from several given options. The answers for Q3 & 20 are positive numbers. More specifically, answer for Q3 is everyday expense on

transportation; answer for Q20 is family size; answer for Q18 is date of birth; answer for Q25 is suggestion; and answer for Q26 is email address. The data set has 671 observations and 73 columns: 1 for ID; 68 for answers of questions; 4 for Internet information of visitors. And missing values exist.

3 Data Preprocessing

Remove irrelevant variables.

ID, suggestion, email address, and Internet information of visitors are removed.

Detect errors.

For single choice question, answer should be restricted in given range. And some answers are obviously wrong, for example age = 200 and family size = 100. So age range is set into [10, 110] and range of family size is set into [0, 20]. All incorrect answers are treated as missing values.

Re-format variables.

- Date of birth (DOB) is originally in the form of "YYYY-MM-DD" as a factor value with 468 levels and converted into numeric value for age.
- Q4_1 - Q4_8: rating for transportation condition ratings for 8 different conditions are converted into a rating with range [8, 37].
- Q5_1 - Q5_12: rating for transportation problems ratings for 12 different problems are converted into a rating with range [12, 72].
- Q15: the results of accident history are converted into a single variable that denotes total loss of all the historical accidents. It is a factor variable with 9 levels.

Create new meaningful variable by combining variables.

- Merge results from Q4 and Q5 by using appropriate weights and create one single variable as overall satisfaction rating for transportation system. The overall rating variable after transformation ranges [0, 100].

- Q16: there are 22 sub-questions about driving habits, including 10 identified as good habits, 10 identified as bad habits and 2 undetermined. A new variable that measures the score of driving habits is created with the range of [0,100].

4 Modeling and Assessment Methods

Besides linear and generalized linear models (LM and GLM), generalized additive models (GAM), tree and neural network (NNET) are also employed for the study. The introduction of these models can be found in many books and will not be included here.

The model assessment is conducted in the following way.

- "Hold-out" method is used to dividing data set into training and testing sets.
- Cross-validation and bootstrapping is used for some cases.
- For NNET, the procedure is repeated many times to obtain average prediction.
- Stepwise variable selection is used for LM and GLM.
- The mean of squared errors of prediction (MSE) and R^2 are used to for evaluating regression problems and the error rate is calculated for classification problems.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.1}$$

$$R^2 = \text{cor}(\boldsymbol{y}, \hat{\boldsymbol{y}})^2 \tag{4.2}$$

5 Implementation and Results

5.1 Case I

Relationship : satisfaction rating ~ time + expense + gender + age + education + family size + income

Response variable : treated as numerical.

Explanatory variables: : "gender" is treated as factor variable; variables - "time", "education" - are well ordered and treated as numerical as well as other variables.

Detail : 466 observations are used and divided into training/testing sets with the ratio 70%/30%.

Modeling : see the attached code for the details of models.

The results are summarized in table 5.1. We can see that:

- tree produces the best prediction for training data but not for the testing data which suggests the "overfitting". This is exactly one of the drawbacks for regression tree. However, by pruning the over-branched tree, the "overfitting" problem is significantly reduced, and actually the pruned tree is one of the best models that perform well on testing data.

- the interaction terms in LM indeed increase the prediction power for training data, but they actually reduce the power for testing data. This is a typical phenomena of "bias and variance tradeoff": the complicated model may provide less bias (for training data) but at the meantime results in higher variance (which may lead to worse prediction for testing data). The simpler models are usually more robust.

- GAM performs well for both training and testing data due to the flexibility of the model.

- NNET acts in similar way as un-pruned tree does – "overfitting" the training data.

- GLMs completely fail on this data set because the assumed distribution for response variable is not close to the true distribution.

Table 5.1: Assessment for Case I

	Training		Testing	
Model	MSE	R^2	MSE	R^2
LM	173.24	0.188	227.46	0.094
LM (SW)	173.33	0.188	227.71	0.093
LM (IT)	157.60	0.262	234.40	0.102
LM (IT+SW)	160.89	0.246	243.09	0.077
GLM	1395.56	0.187	1460.65	0.095
GLM (SW)	1395.56	0.187	1460.65	0.095
GLM (IT)	1395.18	0.259	1459.05	0.101
GLM (IT+SW)	1395.18	0.259	1459.05	0.101
GAM	153.19	0.283	221.41	0.112
TREE	129.77	0.392	270.36	0.054
TREE (prune)	161.69	0.242	231.35	0.088
NNET*	157.73	0.266	263.96	0.045

SW: stepwise; IT: interaction terms

5.2 CASE II

Relationship : driving habits ~ gender + age + education + income

Response variable : treated as numerical.

Explanatory variables: : "gender" is treated as factor variable; "education" is well ordered and treated as numerical as well as other variables.

Detail : 267 observations are used and divided into training/testing sets with the ratio 70%/30%.

Modeling : see the attached code for the details of models.

The results are summarized in table 5.2. In this case, GLMs still cannot fit the data set well because the distribution of the response variable is not close to common distributions, such as Poisson, Gamma, and hard to be identified (see the code for more details). The results of assessment for other models reveal similar patterns we have seen in previous case, except that GAM isn't a good fit for this data set because none of the predictor is significant for the model (see the following testing output).

Table 5.2: Assessment for Case II

	Training		Testing	
Model	MSE	R^2	MSE	R^2
LM	436.95	0.021	521.91	0.059
LM (SW)	441.10	0.012	535.21	0.032
LM (IT)	420.06	0.059	547.51	0.014
GAM	416.95	0.069	718.46	0.022
TREE	363.19	0.187	620.46	0.000
TREE (prune)	432.27	0.032	555.78	0.005
NNET*	343.30	0.299	582.20	0.007

SW: stepwise; IT: interaction terms

```
Call: gam(formula = dhabit ~ lo(age) + lo(edu) + lo(income) + gender,
    data = d2.train)
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-55.180 -12.781   1.905  16.063  42.182

(Dispersion Parameter for gaussian family taken to be 453.8179)

    Null Deviance: 83044.66 on 185 degrees of freedom
Residual Deviance: 77552.35 on 170.8887 degrees of freedom
AIC: 1682.199

Number of Local Scoring Iterations: 2

DF for Terms and F-values for Nonparametric Effects

            Df Npar Df  Npar F  Pr(F)
(Intercept)  1
lo(age)      1     2.6 0.81718 0.4701
lo(edu)      1     4.0 0.22780 0.9225
lo(income)   1     3.5 1.55079 0.1960
gender       1
```

Table 5.3: Assessment for Case III

Model	Training Error Rate	Testing Error Rate
TREE	40.19%	68.09%
TREE (prune)	41.12%	57.45%
TREE (10 CV)	55.14%	53.19%
NNET*	23.36%	6.38%

10 CV: 10-fold cross-validation

5.3 CASE III

Relationship : accident loss ~ driving habits + driving time + frequency

Response variable : ordered categorical.

Explanatory variables: : Variables, "driving time" and "frequency", are well ordered and treated as numerical as well as "driving habits" (numerical score).

Detail : 154 observations are used and divided into training/testing sets with the ratio 70%/30%.

Modeling : see the attached code for the details of models.

The results are summarized in table 5.3. For this data set NNET shows promising compatibility and the error rate is surprisingly only 6.38%. On the other hand, the performance of trees is relatively poor even with pruning. The reason is the size of the data set is small. As tree keeps branching, the number of observations in the node reduce quickly and become too small to provide good fitting. To solve this problem, we use bootstrapping technique to increase the sample size and fit tree models to bootstrapped data. The results shown in table 5.4 suggests that the tree can be better fitted with larger bootstrapped sample size but reaches a threshold when the data is bootstrapped 3000. Further bootstrapping beyond this threshold results in too many redundant observations and start damage the model fitting.

Table 5.4: Tree with 10 CV

Bootstrap	Error Rate
500	32.60%
1000	29.60%
1500	21.07%
2000	13.30%
2500	11.44%
3000	11.10%
3500	11.51%
4000	11.70%

10 CV: 10-fold cross-validation

6 CONCLUSION

"Overfitting" problems for tree and NNET need to be noticed by comparing their prediction power on both training and testing data sets or using cross-validation technique. Complicated models usually are able to provide better fitting for training data but in the meantime result in potential higher variance which may lead to poor prediction for testing data. And one model may be robust or flexible for one data set and relatively poor for another. When using GLMs, identifying the distribution of response variables is very important. Poor assumption on this distribution can dramatically reduce the performance of the model. Large sample size is essential to find a good model. With small sample size, prediction ability of all kinds of models is limited. And bootstrapping can enlarge the sample size and help identifying the potential power of the model.