Study Report on Methods in Data Mining

YUAN Zhe, XU Rendan, XU Donglin, GUO Jiabao

Department of Statistics
United International College, Guangdong 519085, China

DM Group Project, 2013

Outline of topics

This study is to set up a classification model to predict whether income exceeds 50K/yr based upon his characters by using data mining techniques.
The data source is from UCI.

Extraction was done by Barry Becker from the 1994 Census database.

# of Instances	48842	Area	Social	Attribute Characteristics	Categorical, Integer
# of Attributes	14	Date Donated	96-05-01	Missing Values	Yes

The part of samples are listed below:

age	workclass	fnlwgt	education	education_num	marital_status	income
64	Private	66634	Bachelors	13	Divorced	1
55	Private	327589	HS-grad	9	Divorced	0
50	Private	104729	HS-grad	9	Divorced	0
39	Private	32146	Some-college	10	Never-married	0
22	Private	109815	Some-college	10	Never-married	0
38	Private	188503	Some-college	10	Never-married	0
45	Self-emp-inc	34091	Bachelors	13	Married-civ-spouse	1
42	Self-emp-not-inc	119207	HS-grad	9	Never-married	0
45	Private	301802	Bachelors	13	Married-civ-spouse	1
60	Private	152369	Assoc-voc	11	Married-civ-spouse	0

Listing of attributes:
income: $> 50K$, $\leq 50K$
age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country.

1. Polish missing value: workclass, occupation, native_country
2. Remove ineffective attributes: fnlwgt, education_num
3. Draw 5000 training and 1000 testing sample
4. Discrete continues variables: age, hours_per_week, capital_gain, capital_loss

This is association rules.
Find frequent set.
Pick out the high frequent combinations
filter out the rhs income and lift=1.2
filter out the rhs income and lift=1.2
sort the rules by confi.

This is Decision Tree.

```
tree1=rpart(income .,data=train,method="class",cp=0.005)
printcp(tree1)
plotcp(tree1)
print(tree1)
summary(tree1)
plot(tree1,uniform=TRUE)
text(tree1,use.n=TRUE,all=TRUE, cex=.8)
post(tree1,file=train)
tree2=prune(tree1,cp=0.01)
plot(tree2)
text(tree2)
Prediction=predict(tree2,newdata=test)
Pre=vector()
```

This is Naïve Bayes Classifier.

```
mm=naiveBayes(income .,data=train)
Pre2=predict(mm,test)
Pre2[1:20]
error2=sum(Ori!=Pre2)
errorratio2=error2/1000
```

This is Artificial Neural Network.

```
nn=nnet(income .,data=train,size=2,rang=0.1,maxit=1000)
Pre3=predict(nn, test, type = "class")
error3=sum(Ori!=Pre3)
errorratio3=error3/1000
```

This is Bagging and Boosting.

```
c¡-rbind(train,test)
m¡-dim(c)
income¡-as.factor(c[ ,m[2]])
c¡-data.frame(c[,-m[2]],income)
bagging¡- bagging(income ., data=c[1:5000,] , mfinal=10)
boosting¡- boosting(income ., data=c[1:5000, ], boos=TRUE,
mfinal=10)
pre5¡-predict.bagging(bagging, newdata=c[5001:6000, ])
pre6¡-predict.boosting(boosting, newdata=c[5001:6000, ])
```

Decision Tree takes the least time in modelling.

Method	Dec Tree	NBayes	ANN	Bagging	Boosting
Timing	0.395 s	0.063 s	10.138 s	5.168 s	5.635 s

Boosting has the least error rate.

Method	Dec Tree	NBayes	ANN	Bagging	Boosting
Timing	18%	20.5%	18.2%	18%	17.2%

1. Data mining is an efficient tool in identity the pattern in data.
2. Evidently, it takes significant time in modelling which require better algorithm in dealing with big data.
3. The attributes of the data are all categorical data which is a weak point.
4. The general error rate is significant so that it need to be improved.

Computing Environment

COMPUTER

OS	OS X 10.8.3 (12D78)
Processor	2.26 GHz Intel Core 2 Duo
Memory	5 Gb 1067 Mhz DDR3

R PACKAGES

R	3.0.0
class	7.3-7
e1071	1.6-1
rpart	4.1-1
nnet	7.3-6
Matrix	1.0-12
lattice	0.20-15
arules	1.0-13
adabag	3.1

Bibliography

1. C L Blake, C J Merz. *UCI repository of machine learning databases* University of California, Irvine, Department of Information and Computer Sciences. 1998