Study Report on Methods in Data Mining

JJ, Rendan, Tutu, Guabao

Department of Statistics
United International College, Guangdong 519085, China

DM Group Project, 2013

Outline of topics

Goal

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.

Data Preview

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.

This is association rules.

This is Decision Tree.

This is Naive Bayes Classifiers.

This is Artificial Neural Network.

This is Bagging and Boosting.

This is Error Rate.

This is Efficiency.

This is Robustness.

The limit is packages. Date mining is good.

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

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