Census Income Data Set (1994) classification using Decision Tree

Heng Meng

A11461867

Introduction

In this assignment, I used 1994 Census data set. This data set contains 48842 instances and 14 attributes. The data set is separated into training set and test set. Training set has 32561 instances, and test set has 16281 instances. List of attributes are: age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country.

This report is consist of four sections. In the first section, I analyze the data set in an exploratory fashion to understand basic properties of the data. I try to demonstrate the simple relationship between features visually so that hypothesis can be established.

In second section, I identify the task of predicting whether a person can earn more than 50K a year base on his information. I describe a baseline solution and why it is relevant. Also, I investigate a better algorithm to perform this predictive task.

I next talk about related work for this kind of predictive task.

In the last section, I talk about the result and conclusion.

Exploratory analysis

This data set is consist of 10771 females and 21790 males. The people's age vary from 17 to 90 years old. There are 15 different kinds of occupations, and 16 different levels of education.

Occupation distribution:

Pro-specialty	4140
Craft-repair	4099
Exec-managerial	4066
Adm-clerical	3770
Sales	3650
Other-service	3295
Machine-op-inspct	3295
?	1843
Transport-moving	1597
Handlers-cleaners	1370
Farming-fishing	994
Tech-support	928
Protective-serv	149
Armed-Forces	9

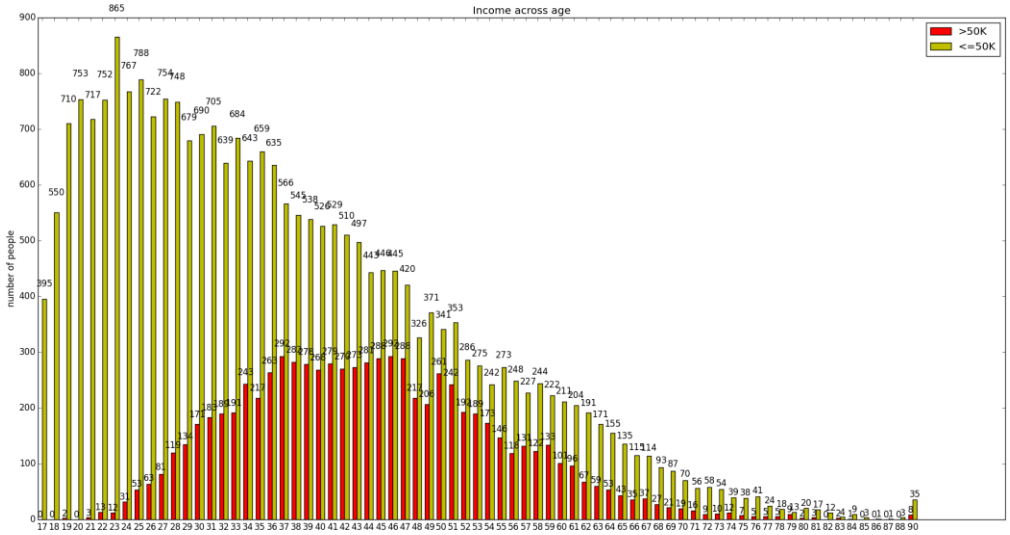
Education levels distribution:

High School Grad	10501
Some-college	7291
Bachelor	5355
Master	1723
Assoc-voc	1382
11^{th}	1175
Assoc-acdm	1382
10^{th}	933
Prof-school	576
9^{th}	514
12^{th}	433
Doctorate	413
$5^{th} – 6^{th}$	333
$1^{st} – 4^{th}$	168
Preschool	51

Work class distribution:

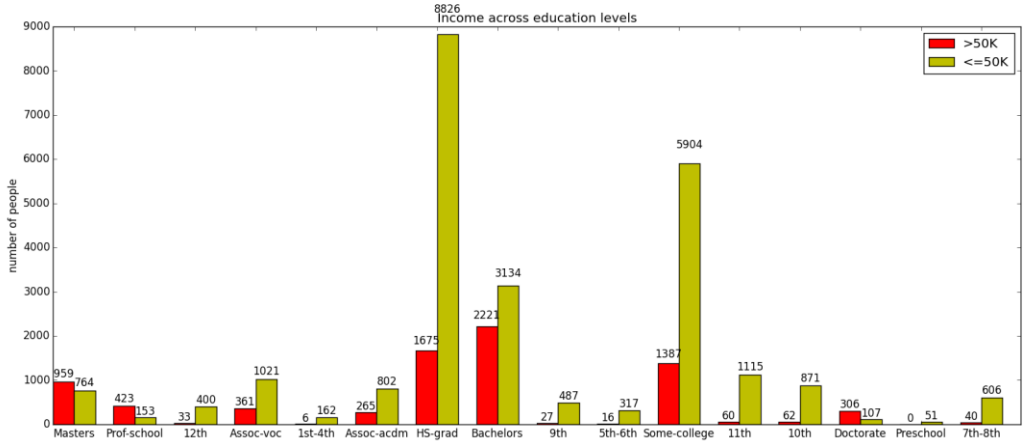
Private	22696
Self-emp-not-inc	2541
Local-gov	2093
?	1836
State-gov	1298
Self-emp-inc	1116
Federal-gov	960
Without-pay	14
Never-worked	7

I am interested in the ratio of number of people who make more than 50K and number of people who make less than 50K. I group people by different attributes. I explore the relationship between age and income.



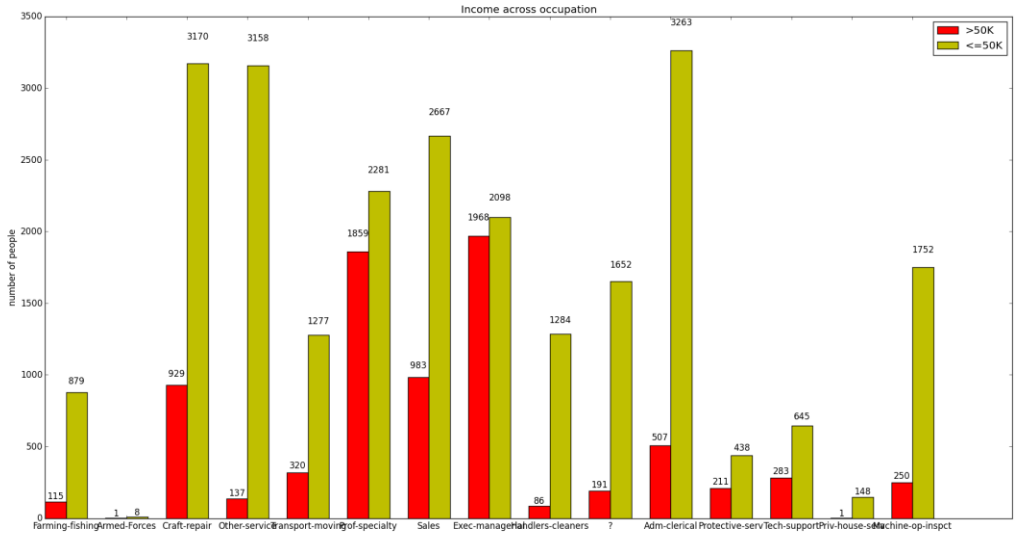
The ratio is very low in age group 17 – 27. The ratio increases as age increases, and the highest ratios appear in age group 40 – 50. The ratio decreases as age increases after age 50.

I also examine the relationship between education level and income.



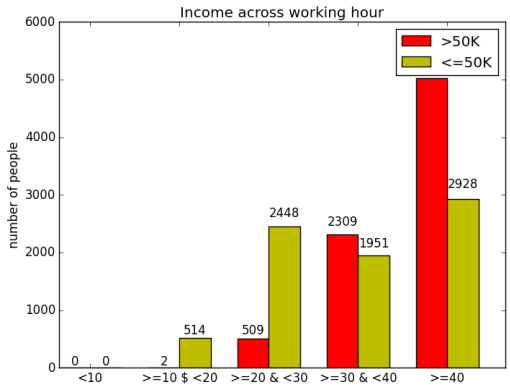
The ratio is very high in group of people who have doctorate degree and group of people who have prof-school degree.

This graph shows income across occupation. In groups pro-specialty and Exec-manage, the ratio is high, while the ratio is low in other-service, handler-cleaner, and Adm-clerical groups.

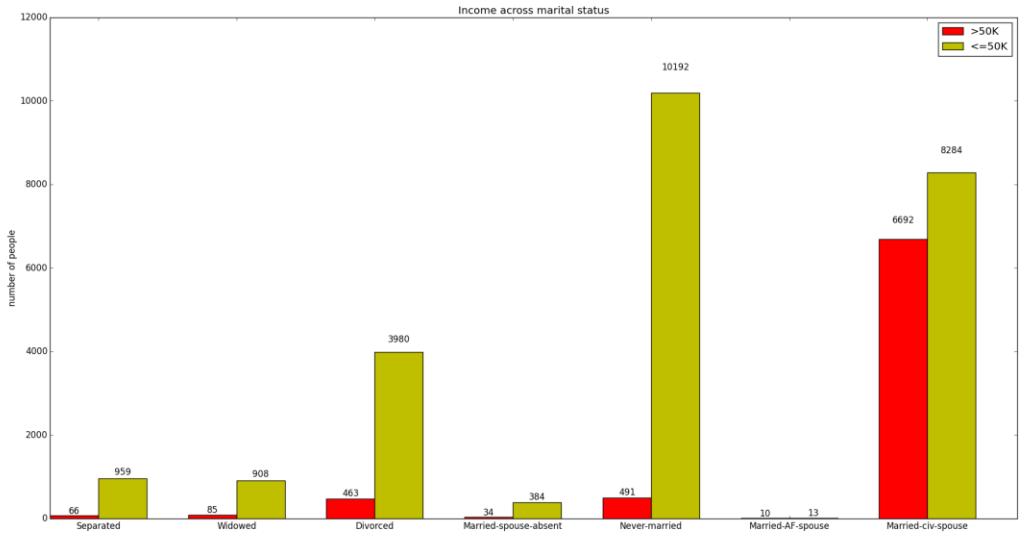


Then, I group people in different ranges of working hours per week. The hours per work is continuous number in data set. I am more interested in discrete feature. It is very obvious

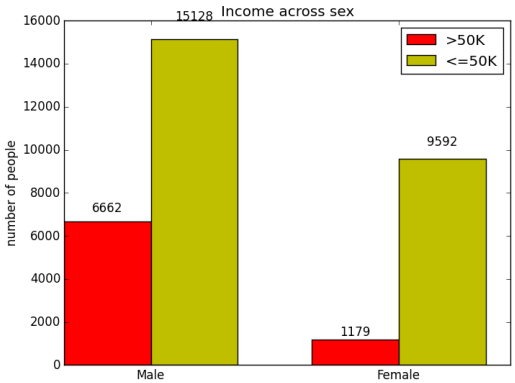
that the ratio is higher in group of people work harder.



This is graph of people grouped by marital status. The ratio is low in all the group except the group married-civ-spouse. We can observe that stable marriage contributes a lot.

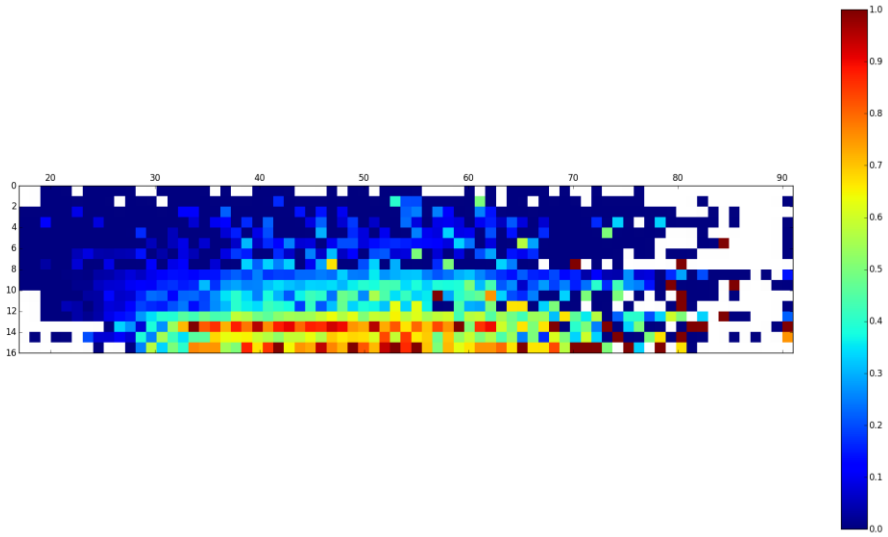


Also, I take look at the distribution difference between males and females.



The ratio is higher in male group than in female group.

Lastly, this graph show the distribution in different combination of age and education.



X-axis represents age, and y-axis represents education level. The larger number represents higher education level. The color of square box represent the ratio. Blue is lowest, and red is highest. White is missing combination.

Identifying a Predictive Task

The predictive task could be predicting whether a person can make 50K a year base on his/her information. Probability of making more than 50K is 23.93%. In previous section, I examine many relationship between income and different features. We can tell that there are many strong indicator that can help us determine whether a person can earn more than 50K a year. For example, people who have high education usually make more than people who have low education. People who work hard make more than people who work less hard. According to those observation, the base line solution can be as following:

If (education == 'doctorate' or education == 'pro-school')

Predicts yes

Else If (working hour >= 40)

Predicts yes

Else

Predicts no

This base line solution actually perform pretty well already. The error rate is 0.37215. However, this native predictor doesn't utilize other useful feature. Also, the threshold are manually tuned. I need better way to utilize other feature and learn threshold.

I use ID3 Decision Tree algorithm for this task. But it is hard to apply this algorithm directly on the data set. There are couple of problems to make it difficult. Many feature are discrete, so threshold doesn't make sense on those features. Also, many features are not numbers. I need to process the data so that the algorithm can be applied on the data set.

I need to replace those non-numeric features with numbers. If I arbitrarily replace features with number I will create many internal fragments. For example, making cut at anywhere might give me same information gain, then the tree will become very complicated or very random. To prevent this happen, replace the features with informational numbers. We can use the statistics found in previous section. Replacing the feature with its corresponding ratio can give me meaningful number rather than arbitrary number.

Related Works

The paper by Ron Kohavi [3] talks about a modified version of ID3 Decision Tree. The new algorithm is called NBTree, which induces a hybrid of decision-tree classifiers and Naïve-Bayes classifiers. The NBTree nodes contain univariate splits as regular decision-tree, but the leaves contain Naïve-Bayesian classifiers.

The paper by Jinyan Li [2] introduces a new algorithm doesn't use distance as measurement, but use frequency of an instance's subsets and the frequency-changing rate of the subsets among training classes to perform both knowledge discovery and classification tasks.

The work by Dennis P. Groth [1] talks about the use of entropy for visualizing database structure. Visualizing entropy of a relation provides a global perspective on the distribution of values and helps to identify areas within the relation where interesting relationships may be discovered.

Conclusion and Results

With the ID3 Decision Tree algorithm and the reprocessing I describes in second section, I am able to get error rate 0.1754. Comparing to the error rate 0.37215 from the baseline solution, my algorithm improved by 0.19675. There is error rate list of other algorithm running on this data set.

Algorithm	Error rate
C4.5	0.1554
C4.5-auto	0.1446
C4.5 rules	0.1494
Voted ID3 (0.6)	0.1564
Voted ID3 (0.8)	0.1647

T2	0.1684
1R	0.1954
NBTree	0.141
CN2	0.16
FSS Naïve Bayes	0.1405
Nearest-neighbor (1)	0.2142
Nearest-neighbor (3)	0.2035

Comparing to the list of algorithms, my algorithm does a reasonable job. In previous section, I talk about the NBTree algorithm. It is similar as my algorithm. The difference is that leaves in NBTree are Naïve Bayes classifiers, while I use the original Decision Tree algorithm with preprocessed features.

There is almost 20% accuracy improvement of my algorithm comparing to baseline solution. From first section, I learn that there are many good features can be used to classify, but it is not clear how to use them. ID3 algorithm provides a way to find good features and thresholds by computing information gains. Also, preprocessing data set plays an important role in my algorithm, since many features in original data set are not in proper forms.

Reference

[1] Dennis P. Groth and Edward L. Robertson. An Entropy-based Approach to Visualizing Database Structure. VDB. 2002.

[2] Jinyan Li and Guozhu Dong and Kotagiri Ramamohanarao and Limsoon Wong. DeEPs: A New Instance-based Discovery and Classification System. Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases. 2001.

[3] Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996