A decision tree for predicting the income individuals from U.S. Census data

Mark Patrick White
CSIT 599G Computational Data Mining
Hood College, Computer Science Department
Frederick, Maryland, USA
mpw6@hood.edu

A decision tree for predicting the income individuals from U.S. Census data

- Project Overview
- Introduction and Background
- Approach and Implementation
- Results
- Future Work
- Conclusions
- Questions and Discussion

Project Overview

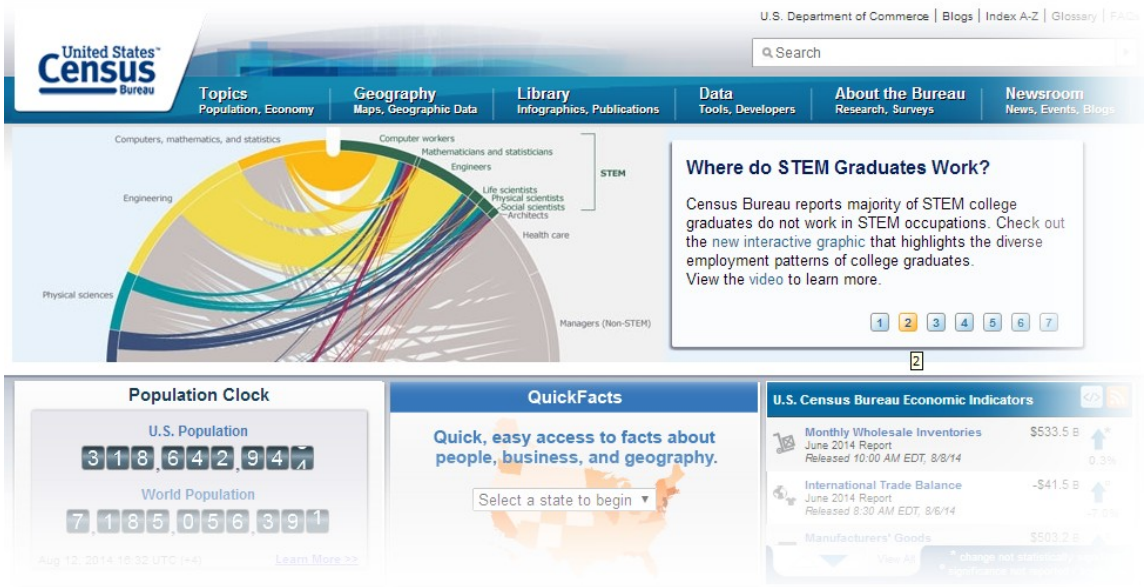
- The intent of this project is to use a standard data mining technique (classification with a decision tree algorithm) on a collection of U.S. Census data in order to predict the income range of future census entries.
- The goal of the project is to make accurate predictions on whether a person makes either up to or more than $50,000 annually, based upon certain attributes associated with that person's census information.

Project Overview

- U.S. Census Bureau Data (census.gov)
- Decision Tree can be used to predict an individual's income using a subset of U.S. Census data
- Modified ID3 Decision Tree
- More data improves results, but also can complicate results

Introduction and Background

- U.S. Census Bureau
- http://www.census.gov/



Introduction and Background

- Part of the U.S. Department of Commerce
- Primary mission is the U.S. Census every ten years, which allocates the seats of the U.S. House of Representatives to the states based on their population.
- In addition to the decennial census, the Census Bureau continually conducts dozens of other censuses and surveys, including the American Community Survey, the U.S. Economic Census, and the Current Population Survey.
- Bureau censuses and surveys are used by the federal government help assign budgetary allocations to various agencies and jurisdictions.

Approach and Methodology

Approach

1. Preprocess a sample of U.S. Census data, collected from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Census+Income)
2. Use ID3 Decision Tree Algorithm with training set to generate a decision tree
3. Process test set using driver application running the decision tree logic to show number of hits/misses for decision tree predictions

Approach and Methodology

ID3 Algorithm Summary

1. Calculate the entropy of every attribute using the data set
2. Split the set into subsets using the attribute for which entropy is minimum (or, equivalently, information gain is maximum)
3. Make a decision tree node containing that attribute
4. Recurse on subsets using remaining attributes

Approach and Methodology

UCI Dataset Attributes

- **age**: continuous.
- **workclass**: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- **fnlwgt**: continuous. **(???)**
- **education**: 16 distinct education attribute classes.
- **education-num**: continuous.
- **marital-status**: 7 distinct marital status attribute classes.
- **occupation**: 14 distinct occupation attribute classes.
- **relationship**: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- **race**: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- **sex**: Female, Male.
- **capital-gain**: continuous. **(???)**
- **capital-loss**: continuous. **(???)**
- **hours-per-week**: continuous.
- **native-country**: 41 distinct native country attribute classes.

Approach and Methodology

Test Execution Dataset Attributes

- **age_class**: 5 distinct age attribute classes (U21, U30, U40, U55, 55PLUS).
- **workclass**: 8 distinct workclass attribute classes.
- **education**: 16 distinct education attribute classes.
- **marital_status**: 7 distinct marital status attribute classes.
- **occupation**: 14 distinct occupation attribute classes.
- **relationship**: 6 distinct relationship attribute classes.
- **race**: 5 distinct race attribute classes.
- **sex**: Female, Male.
- **hours_class**: 5 distinct hours-per-week attribute classes (UPTO20, UPTO40, UPTO60, UPTO80, 80PLUS).
- **native_country**: 41 distinct native country attribute classes.

Target Test Attribute

- **income_class**: <=50K, >50K

Approach and Methodology

Test Execution Dataset Attributes

- **age_class**: 5 distinct age attribute classes (U21, U30, U40, U55, 55PLUS).
- **workclass**: 8 distinct workclass attribute classes.
- **education**: 16 distinct education attribute classes.
- **marital_status**: 7 distinct marital status attribute classes.
- ~~**occupation**: 14 distinct occupation attribute classes.~~
- **relationship**: 6 distinct relationship attribute classes.
- **race**: 5 distinct race attribute classes.
- **sex**: Female, Male.
- **hours_class**: 5 distinct hours-per-week attribute classes (UPTO20, UPTO40, UPTO60, UPTO80, 80PLUS).
- ~~**native_country**: 41 distinct native country attribute classes.~~

Target Test Attribute

- **income_class**: <=50K, >50K

Results

- Trouble in Prediction Paradise
- Resulting Decision Tree was over 118,000 lines of code
- Java compiler can only handle up to 64K of code
- Sequential scale back of learning data was required

Results

1. Learning set of 3500 records with 10 attributes (118000+ lines of code)
2. Learning set of 500 records with 10 attributes (35000+ lines of code)
3. Learning set of 3500 records with 8 attributes (5000+ lines of code)
4. **Learning set of 2000 records with 8 attributes** (~4000 lines of code)
5. **Learning set of 500 records with 8 attributes** (~2000 lines of code)

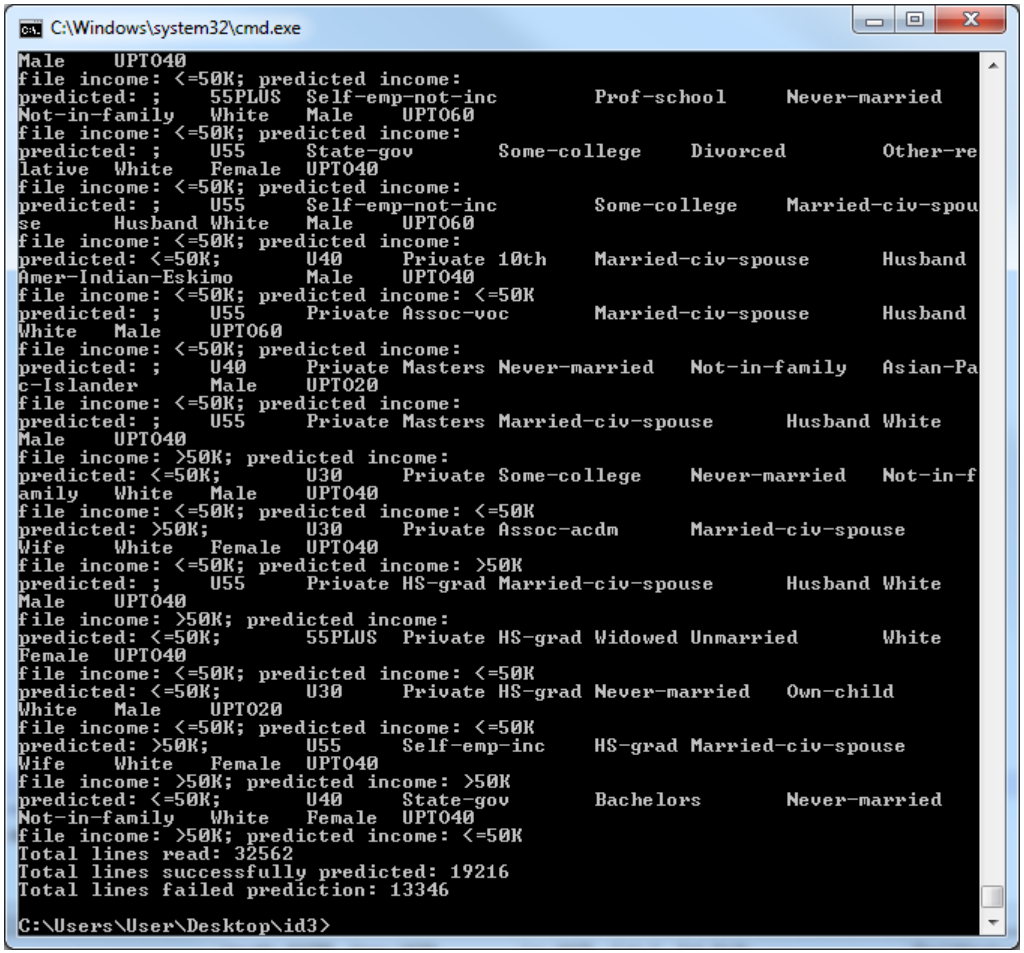
A decision tree for predicting the income individuals from U.S. Census data

Results

- **Learning set of 2000 records with 8 attributes** (~4000 lines of code)

- **Total lines read: 32562**
- **Total lines successfully predicted: 19216**
- **Total lines failed prediction: 13346**

- **Success (%): 59.0%**



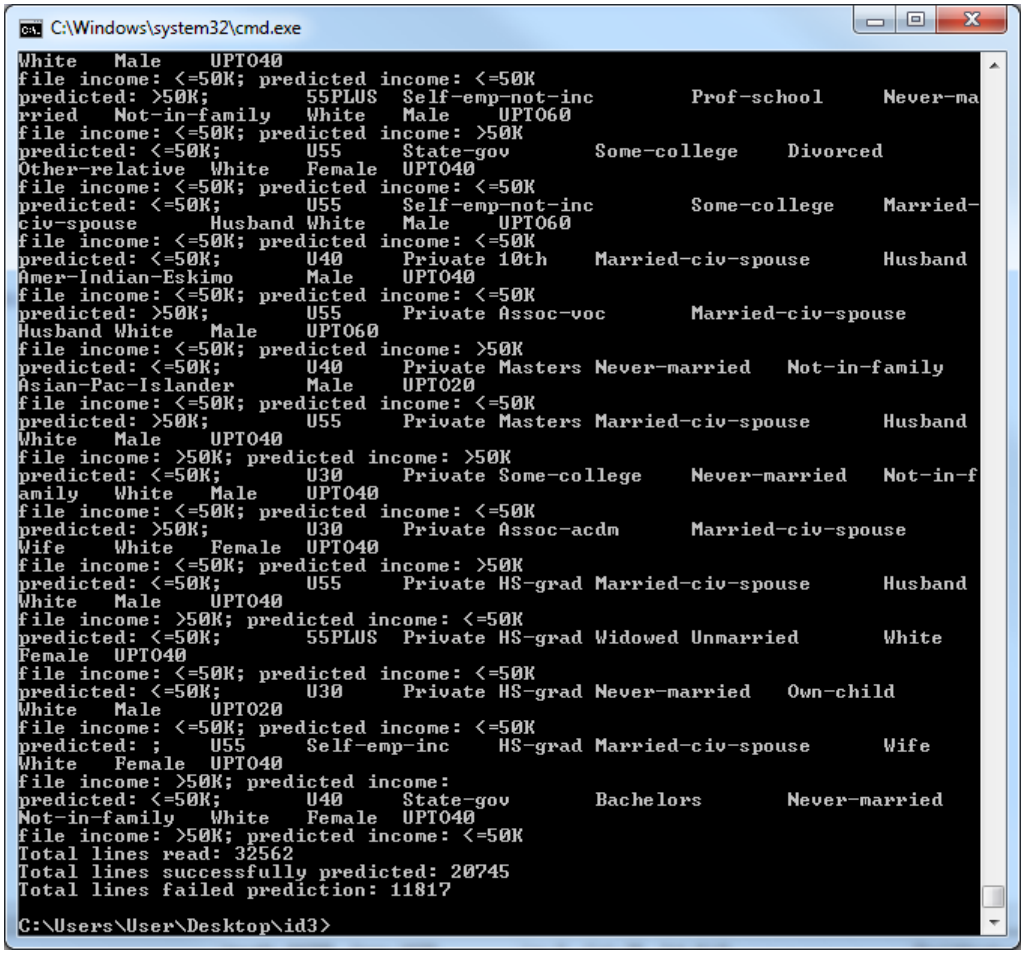
A decision tree for predicting the income individuals from U.S. Census data

Results

- **Learning set of 500 records with 8 attributes** (~2000 lines of code)

- **Total lines read: 32562**
- **Total lines successfully predicted: 20745**
- **Total lines failed prediction: 11817**

- **Success (%): 63.7%**



Results

- **Learning set of 2000 records with 8 attributes: 59.0%**
- **Learning set of 500 records with 8 attributes: 63.7%**

- What's going on here?

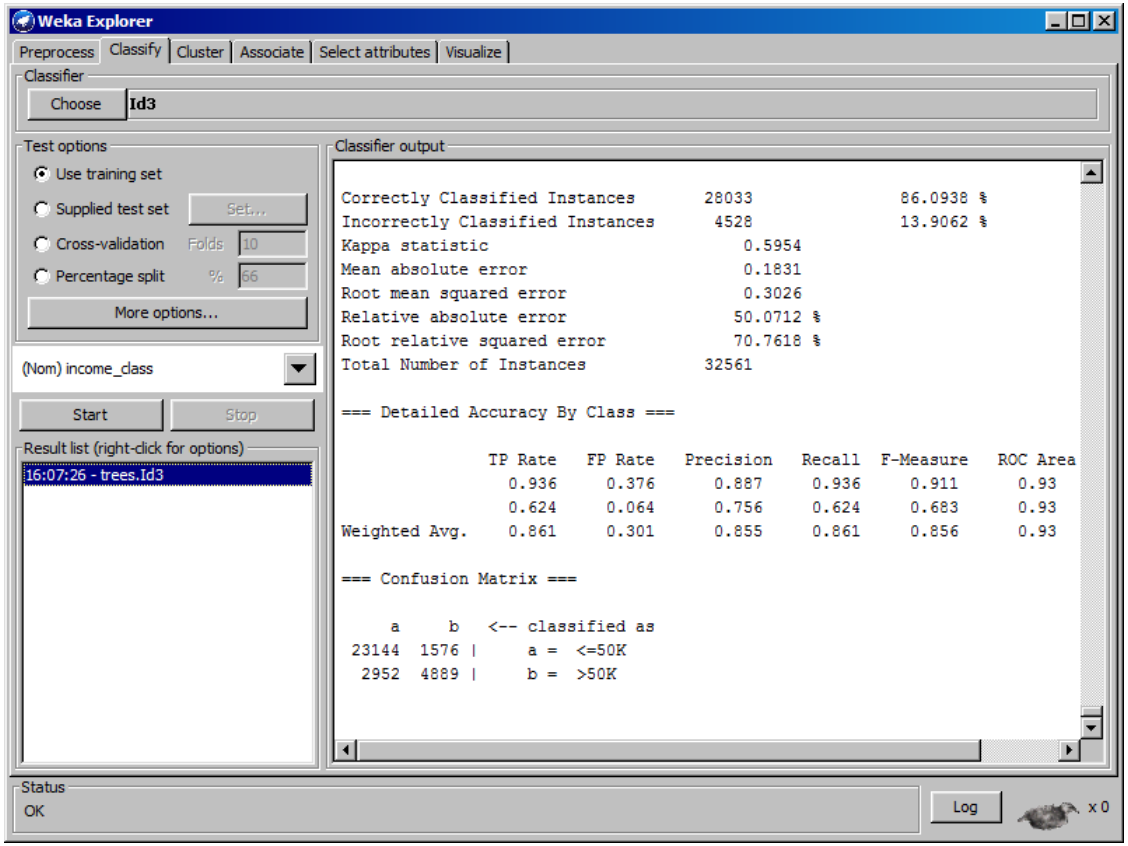
A decision tree for predicting the income individuals from U.S. Census data

Results

- **Weka (ID3 with learning set of all 32562 records)**

- **Total lines read: 32562**
- **Total lines successfully predicted: 28033**
- **Total lines failed prediction: 4528**

- **Success (%): 86.1%**



Future Work

- Use larger subset of U.S. Census data
- Include more attributes into the datasets
- Use different classification algorithms
- Streamline ID3 decision tree generation code

Conclusions

- Decision Tree prediction is heavily dependent upon the dataset (record variety, attributes).
- Decision Tree prediction is heavily dependent upon the quality of the decision tree generation algorithm.
- ID3 Decision Tree Classification is a fairly effective means of classifying/predicting data.

References

- Saini, R., P. Singh. “Classification using decision trees,” Internet: http://iasri.res.in/ebook/win_school_aa/notes/Decision_tree.pdf. [July 18, 2014].
- Bache, K., Lichman, M., “Census Income Data Set”. Internet: https://archive.ics.uci.edu/ml/datasets/Census+Income. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. [July 18, 2014].
- Wikipedia. “ID3 algorithm,” Internet: http://en.wikipedia.org/wiki/ID3_algorithm. [July 18, 2014].
- Antonov, A., "Classification and association rules for census income data", Internet: http://mathematicaforprediction.wordpress.com/2014/03/30/classification-and-association-rules-for-census-income-data/. Mathematica for prediction algorithms: Using Mathematica implementations of machine learning algorithms [http://mathematicaforprediction.wordpress.com]. (March 30, 2014).

Questions?

A decision tree for predicting the income individuals from U.S. Census data

Thank you!

Mark Patrick White
CSIT 599G Computational Data Mining
Hood College, Computer Science Department
Frederick, Maryland, USA
mpw6@hood.edu