Predicting income category using decision tree classification techniques on a multivariate collection of U.S. Census data

Project Proposal

Mark Patrick White
CSIT 599G Computational Data Mining
Hood College, Computer Science Department
Frederick, Maryland 21701
mpw6@hood.edu

I. INTRODUCTION

Given the large amount of census data collected every ten years by the United States Census Bureau, demographic information for the American population is readily available and highly detailed. Taking into consideration the depth of the demographic data available, certain patterns and trends in income, gender, and other identifying features of U.S. citizens should be observable under a certain level of categorical scrutiny. The intent of this project is to use typical data mining techniques (most likely classification with a decision tree algorithm) on a large 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.

II. DATA SET

The data set consists of over 65,000 instances of census data, each instance representing a single citizen census entry. This data was collected by Barry Becker (of Silicon Graphics' Data Mining and Visualization team) from the 1994 U.S. Census database, from which a clean subset of data was extracted (filtered for citizens over sixteen that work a non-zero amount of hours-per-week and make over $100 annually). The data was subsequently donated in 1996 to the UCI Machine Learning Repository, from which the data set was taken [2]. The data set consists of fourteen attributes to represent each instance, plus the classifier value (the annual income category) to be used in training the algorithm. The attribute fields include personal attribute types like the age of the citizen, their education level and a numeric representation of that education level, the citizen's marital status, family relationship (Wife, Husband, Not-in-family, etc.), as well as their race, national origin, and sex. Also, there are work-related attributes like "work class" (whether they are privately-employed, a government employee, or self-employed), capital gain and loss, hours-per-week worked; as well as the income classification category (>50K, <=50K). The attributes are a mix of types: categorical with multiple levels, binary, and numerical.

III. METHODOLOGY

The methodology intended to be employed will involve a variant of a decision tree algorithm, modified to accept several attributes in the classification attempt [1]. This classification algorithm should determine the expected salary category from a set of data preprocessed to eliminate data records where over a third of the attributes used by the decision tree are missing.

IV. REFERENCES

[1] Saini, R., P. Singh. "Classification using decision trees," Internet: http://iasri.res.in/ebook/win_school_aa/notes/Decision_tree.pdf. [July 18, 2014].

[2] 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].