The Design and Implementation of Decicion Trees for Career Guidance

Casey Barrett

April 14, 2005

1 Abstract

This research project will be an investigation into the design and implementation of various decision trees for career guidance. A decision tree takes into account some sort of situation outlined by a group of parameters and outputs a Boolean decision to the situation. This project will take into account many aspects associated with decision trees including database building, searching and sorting, and algorithms for accessing data.

My project utilizes numerous decision trees in an effort to serve as a tool for career guidance for young adults. A user will fill out a form of specified fields that will then be analyzed by the group of decision trees until a field of study/occupation is given to the user as the outcome. This group of decision trees will be built through database building techniques.

I have utilized extensive research from various websites that offer expertise in a number of areas. I used tutorial websites that had examples and explanations of decision trees generated by the C4.5 program. I have also tried to read various research papers that deal with decision trees. Although I have yet to find one that remotely relates to career guidance, I feel that my understanding of decision trees has increased as a result of these research papers.

2 Introduction

This project will utilize numerous decision trees to assist young people by helping them focus on their interests and what career paths may coincide with those interests.

3 Background/Implementation

A decision tree is a graphical representation of the decision analysis process. This type of tool consists of some sort of input, whether it is a situation or an object. This input is then sent through a set of parameters, or rules, and

eventually the tree gives a Boolean output. There are many different types of parameters that can be used. Or in other words, many different types of parameter cases can be used. These cases can include numerical data, simple yes/no answers, or word answers, such as hair color (black, brown, or blonde). Each parameter will have a specified set of cases that correspond to the parameter.

To help me better understand decision trees and how to build them, I have utilized Professor Ross Quinlans revolutionary decision tree generator programs, the C4.5. What these programs do is take into account large sets of data (from properly written database files) and looks for correlations within the sets. It then uses these correlations to build a decision tree that follows the 'rules' outlined by the correlations. In an auxiliary program, the C4.5rules program, the 'rules' for the tree generated can be displayed.

This program gave me a better understanding of how decision trees are built. I wrote my own database files that could be read by the C4.5 program. I learned the proper syntax for these database files, which I would later utilize for career guidance. For each database, two separate kinds of files are needed, the .names file, which outlines each parameter and the appropriate cases as well as the end cases, and also the .data file, which consists of singular entries. Each entry in the database properly fills up each parameter outlined in the .names file.

I also had to research techniques for career guidance. The most rudimentary of these techniques was to have a user fill out a list of field data and then compare the user's answers to those of a highly comprehensive database. Some of these questionnaire-type devices included many different occupational fields, such as art/music, engineering, writing/journalism, and social services. For my project, I decided to start out with two very distinct fields that would be a good way to acclimate myself for making a career guidance program. These two fields would be the liberal arts field and the sciences field.

For career guidance, I separated the decision trees that I would need into three different and distinct trees. The first one is a tree designed to help a user decide whether they should focus on either of the broad intellectual categories of the sciences or the liberal arts. This decision will be decided based a series of fields the user fills out in a separate C++ program. Some of these fields include the user's grades in their current English and Math class, if the user is in a science club, the number of computers that the user owns, and the number of plays the user has participated in during the last year. These starting questions are somewhat broad because this is the first preset tree that the user will be compared to.

The next step was to develop a user input program that prompts the user questions related to career guidance. The program then takes the answers and writes them to a .data file. I wrote this program in C++. Likewise, I needed to use the fstream.h library in order to gain access to the classes ifstream and ofstream. I then created the questions test file that became the ifstream that the program will read. This program contained four functions, better outlined in the Iteration Report: Third Quarter. Currently, the only acceptable inputs for this program are strings, however, during the fourth quarter, I will expand the inputs to include classes char, int, and double.

The user input is first compared to a database of other people's answers and the decision that each of the other people decided upon. Since I have not had enough users, the databases are fictitious data that I have engineered in order to fit my ideal decision tree for this progression. The data from this database helps comprise a decision tree that the user's answers will be run through. Since there are three different decision trees, three separate ideal databases will be made, one for the broad first test, and then one each for the science fields and the liberal arts field.

After a first decision is made, it will be relayed back to the user. If the user wishes to continue, he or she will be given another set of parameters that will be more in depth in either the liberal arts or science fields. These trees, which are still in progress, will help the user compare their answers against another database corresponding to their broad interests (liberal arts or sciences).

When compared to these more specialized databases, a more focused decision was sent back to the user. This is because the user's responses will be matched up with the decision tree that the databases helped generate and the decisions that the databases output. Right now, I am working on the algorithm that compares the user input answers to the decision tree. This algorithm will produce some sort of numerical correlation between the sets of answers. The higher the correlation number is with respect to that specific path of the decision tree, then the more likely the program will output an answer similar to that of the higher correlation. In the end, a type of occupation or a field of study will be output to the user based on the correlations.

Also, I have trained extensively in the ways of the Reverse Game. Without possessing a shred of natural Reverse Game ability, I have worked my way up from the ranks of novice to a somewhat respectable player capable of defeating each and every of the 20 progressively challenging levels.

4 Conclusions

5 References

Decision Trees: A subtopic of machine learning. http://www.aaai.org/AITopics/html/trees.html

Quinlan, Ross. (n.d.) Ross Quinlan Personal Page. http://www.rulequest.com/Personal/

C4.5 Tutorial.
http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html

Career Key: Job Interests. http://www.careerkey.org/cgi-bin/ck.pl?action=choices

Hettich, Scott. (n.d.) UCI Machine Learning Repository
http://www.ics.uci.edu/~mlearn/MLRepository.html

Nilsson, Nils. (May 10, 2004). Introduction to Machine Learning
http://robotics.stanford.edu/people/nilsson/mlbook.html