Voter Classification

Richard Beindorff

Executive Summary

With every major election, many groups and individuals attempt to predict which political candidate someone will vote for. For example, political candidates need to be able to determine groups that would normally vote favorably towards them. They also need to know which groups would traditionally vote against them so they may be able to alter their campaign to attract those voters. Major media outlets also attempt to predict election outcomes on a state by state basis. With this need to predict which candidate a person will vote for, a reliable method must be found. Using individual biographical information collected from potential voters, for example age, education, and gender, a method of predicting which candidate an individual will vote for can be found.

A set of preexisting voter data containing "Age", "Years of Education", "Number of Degrees" earned, "Gender", and the chosen candidate, One, Two, or Three, will be used. The data set also includes another attribute, "Age Category" which categorizes the "Age" attribute in to four different categories, 1, 2, 3, and 4. To begin, different attributes were compared to determine the correlation of the attributes with the chosen candidate from a given data set. The data set was then run through Weka's implementation of the C4.5 algorithm to create a decision tree. It was found that removing certain attributes as well as changing the input parameters gave a final highest accuracy rate of only 58.7% along with a very large tree. Using different input parameters ultimately yielded a significantly smaller tree but with an accuracy rate of only 56.5%. Overall however, the results were determined to be inconclusive as the accuracy rate

could not be improved beyond 60%. Instead it is recommended that more information about the voters be collected in order to better determine which candidate an individual will vote for.

Problem Description

Using individual biographical information collected from potential voters, such as age, education, and gender, a method of predicting which party individuals will vote for must be found.

Analysis Technique

The database "voter.xls" contains 1,847 entries. Each entry contains attributes regarding an individual's "Age", "Gender", "Years of Education", "Number of Degrees" earned, and the presidential candidate they voted for. Another attribute that is listed is "Age Category". The "Age Category" attribute is used to convert the continuous "Age" attribute in to one of four discrete categories. Ages 34 and below are listed as being in category 1, ages 35 through 44 are listed as being in category 2, ages 45 through 64 are listed as being in category 3, and ages 65 and above are listed as being in category 4.

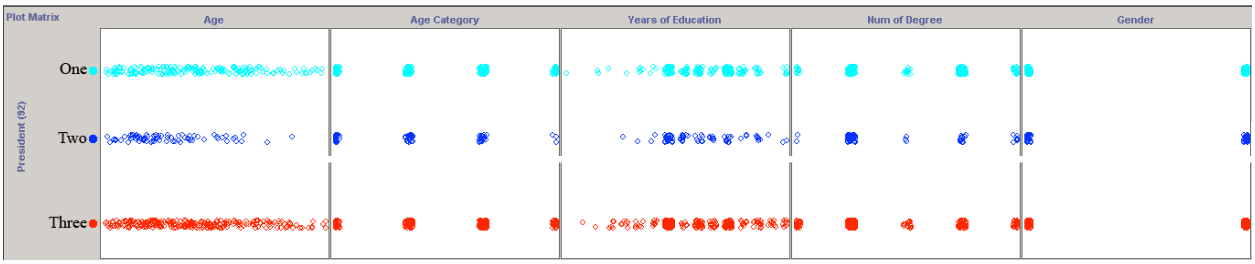
To begin, the correlation coefficient for each attribute was determined. The correlation coefficient is a number which measures the interdependence of two random variables. The correlation coefficient ranges between -1 and +1. Having a value of 1 means the two variables are in perfect correlation with each other. Having a value of -1 means the two variables are in perfect negative correlation with each other. Having a value of 0 means the two variables have no correlation with each other. Using Excel, a table of correlations coefficients were calculated for each attribute versus every other attribute.

Correlation

	Pres (92)	Age	Age Category	Years of Education	Num of Degrees	Gender
Pres (92)	1.0000	0.0297	0.0447	-0.0082	-0.0108	0.1024
Age	0.0297	1.0000	0.9393	-0.2143	-0.2144	0.0126
Age Category	0.0447	0.9393	1.0000	-0.1848	-0.1713	0.0009
Years of Education	-0.0082	-0.2143	-0.1848	1.0000	0.6003	-0.0122
Num of Degrees	-0.0108	-0.2144	-0.1713	0.6003	1.0000	-0.0269
Gender	0.1024	0.0126	0.0009	-0.0122	-0.0269	1.0000

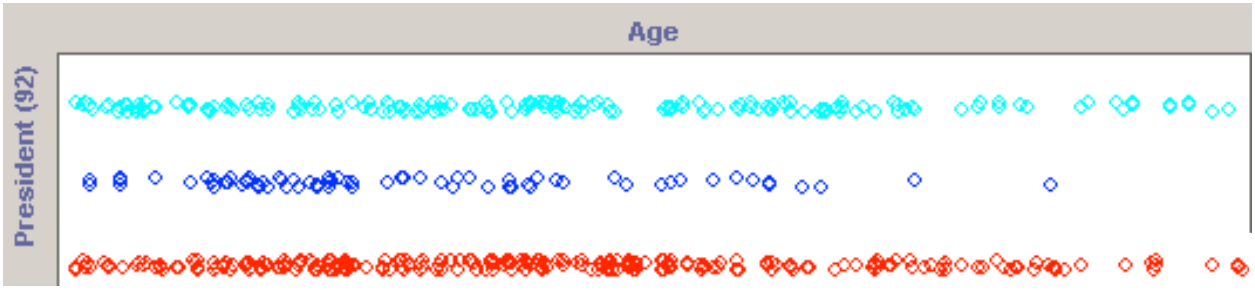
From calculating the correlations of the attributes, it was found that the attributes "Age" and "Age Category" have a correlation of about 0.94 which is the best of all the attributes. However this is to be expected as "Age Category" is just a representation of the "Age Attribute". The second best correlation is between "Years of Education" and "Number of Degrees" with about 0.6. This is expected as well because usually people with a high number of degrees will have more years of education than people with few or zero degrees. Comparing the correlations of "President" to all other attributes gives the highest correlation with "Gender" however it does not give any clear results as they are all fairly close to zero.

Second, the attributes were compared using Weka's Visualize option. After loading the data, Weka is able to create a graph of each attribute versus each other attribute. The "President" attribute was visually compared with the other attributes to determine if any natural distinctions or clusters were formed.



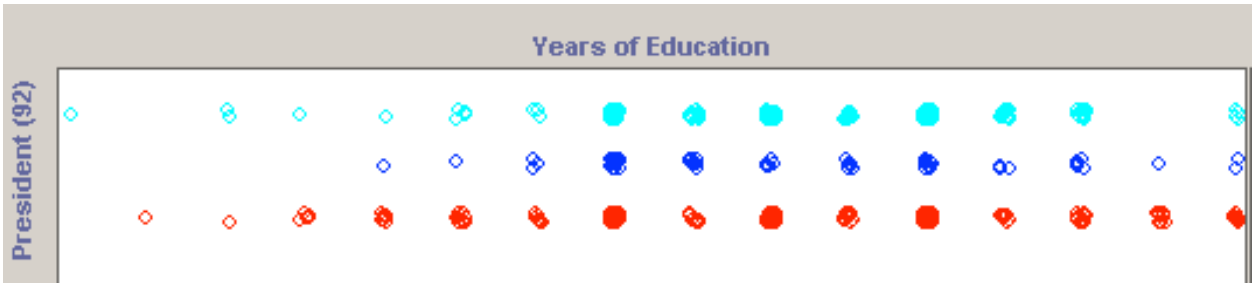
(Weka)

Using information from the first graph, it can be seen that as "Age" increases, voters tend to not vote for candidate Two.



(Weka)

Using information from the third graph, it can be seen that as the "Years of Education" increase, more voters tend to vote for candidate Three. It can also been seen that voters with fewer years of education are not likely to vote for candidate Two.



(Weka)

From these graphs it is also possible to determine that the data does not form any natural distinct groups when compared to the "President" attribute. Because of this, clustering algorithms such as K-Means and K-Nearest Neighbor would most likely not be a good method of classifying voters and were not chosen. Instead, the C4.5 decision tree algorithm will be used to classify voters.

The C4.5 algorithm is able to create a decision tree based on the information gain of different attributes in the data set. Depending on certain inputs, it is also able to "prune" the tree in order to make the tree smaller and easier to work with. For example, a large tree may be able to classify data with a higher accuracy than a small tree however the large size could negatively affect the usage of the tree. With pruning, the tree is able to be shortened. While this could

impact the accuracy, it can also allow it to be implemented much more quickly possibly making it the better option over a larger tree.

Using Weka's built in C4.5 algorithm, called “J48” within the Weka, the entire data set was processed using all attributes with the default settings* while using the entire data set as a training set. However this only gave an accuracy of 54.7374 %. To try to increase this accuracy, the data was sent through the algorithm again; this time removing attributes, one at a time and two at a time, to see if any significant changes were made.

Removed: “Gender”		
Correctly Classified Instances	1005	54.4126 %
Incorrectly Classified Instances	842	45.5874 %

Removed: “Age”		
Correctly Classified Instances	956	51.7596 %
Incorrectly Classified Instances	891	48.2404 %

Removed: “Age Category”		
Correctly Classified Instances	1038	56.1992 %
Incorrectly Classified Instances	809	43.8008 %

Removed: “Years of Education”		
Correctly Classified Instances	965	52.2469 %
Incorrectly Classified Instances	882	47.7531 %

Removed: “Num of Degree”		
Correctly Classified Instances	1044	56.5241 %
Incorrectly Classified Instances	803	43.4759 %

Removed: “Age” and “Age Category”		
Correctly Classified Instances	938	50.7851 %
Incorrectly Classified Instances	909	49.2149 %

Removed: “Age” and “Years of Education”		
Correctly Classified Instances	934	50.5685 %
Incorrectly Classified Instances	913	49.4315 %

Removed: “Age” and “Num of Degree”		
Correctly Classified Instances	972	52.6259 %
Incorrectly Classified Instances	875	47.3741 %

Removed: "Age" and "Gender"

Correctly Classified Instances	915	49.5398 %
Incorrectly Classified Instances	932	50.4602 %

Removed: "Age Category" and "Years of Education"

Correctly Classified Instances	966	52.301 %
Incorrectly Classified Instances	881	47.699 %

Removed: "Age Category" and "Num of Degrees"

Correctly Classified Instances	1035	56.0368 %
Incorrectly Classified Instances	812	43.9632 %

Removed: "Age Category" and "Gender"

Correctly Classified Instances	945	51.164 %
Incorrectly Classified Instances	902	48.836 %

Removed: "Years of Education" and "Num of Degree"

Correctly Classified Instances	1015	54.954 %
Incorrectly Classified Instances	832	45.046 %

Removed: "Years of Education" and "Gender"

Correctly Classified Instances	958	51.8679 %
Incorrectly Classified Instances	889	48.1321 %

Removed: "Num of Degrees" and "Gender"

Correctly Classified Instances	985	53.3297 %
Incorrectly Classified Instances	862	46.6703 %

A maximum of only two attributes at a time were removed as removing more would impact accuracy since more than half of the data set would be removed. It was found that the accuracy rates did not improve much from using all attributes and that the best accuracy occurred when removing "Num of Degrees" resulting in 56.5% accuracy. This highest accuracy also gave a relatively small tree that was a size of 67 and 34 leaves. In order to improve accuracy, different inputs were used. First the "confidence Factor" input was increased from 0.25 to 1.0. The confidence factor is used for pruning with smaller values leading to more pruning of the tree and thus less accuracy(Weka). After adjusting this input, the data was run again with the attribute "Num of Degrees" removed. This yielded an accuracy rate of about 57.9 %, slightly

better than before. However, this new tree was much larger with a size of 139 and 70 leaves. After that, another input was changed, the minimum number of objects per leaf. This input was lowered from the default of 2 to 0. When running the data through again, an accuracy of 58.7 % was found with a tree size of 185 and 95 leaves. While again slightly better, this tree is even larger than the second. With more attempts, the accuracy of the method was never able to exceed 60%.

Assumptions

On closer inspection of the dataset, it was found that two entries under "Years of Education" were set to 98, a number which is highly improbable as both entries list an age under 50 years old. To correct this, the average of the "Years of Education" entries were calculated for the remaining entries and rounded to the nearest whole number, 14. This number was then used to replace both 98 entries for the calculations used.

* The default C4.5, or "J48" in Weka, input values used were:

binarySplits = False
confidenceFactor = 0.25
minNumObj = 2
numFolds = 3
reducedErrorPruning = False
seed = 1
subTreeRaising = True
unpruned = False

Results

While being able to correctly classify 1084 entries out of the total of 1847, a method with 60% or more accuracy was not able to be found. All accuracy results were found to be near 50% accurate which is very inconclusive. When removing the attribute "Num of Degrees" from the data set, this method is able to give a maximum accuracy of 56.5%-58.7% depending on the input values however these changes give much larger trees than using the default input values.

While this method may suffice with less than 60% accuracy, it is recommended that additional voter data be collected in order to better determine which candidate a voter will vote for.

Issues

The main issues were the lack of correlations between the attributes. The attribute "Age Category" could most likely be removed as "Age Category" is just a categorical representation of the "Age" attribute. Another issue was the incorrect entries for "Years of Education". While most likely not having any effect, they could have impacted the results. Other issues were that the attributes for each voter did not seem to have much impact on who they voted for.

Appendices

The citation (Weka) was used to denote pictures or information from the Program Weka. Weka can be found at: http:// www.cs.waikato.ac.nz/ml/weka

The citation that is not a picture is used to reference a help file within the Weka program, specifically in the "Weka Explorer" under the "Classify" tab after choosing the "J48" algorithm and clicking on the numbers beside the "Choose" button to change the input attributes. From there the "More" button was clicked to find the needed information about the different input values.