Lauren Flanakin
C4.5
Math 3210

Executive Summary

This paper shows how the algorithm C4.5 can outperform ID3 by creating decision trees from noisy and missing data. Noisy data is data that makes it hard to create decision trees or create rules where classes are easily defined. Missing data is data where values in attributes are not present, or a blank space. C4.5 was designed to improve on ID3's shortcomings. To test if C4.5 was actually designed to improve on ID3's shortcomings, C4.5 will interpret noisy and missing data to create a decision tree. The percentage of error of the decision tree will demonstrate if C4.5 is an improved algorithm. The datasets had to be formatted in comma delimited for C4.5 to interpret them. The datasets also had to be formatted so there was a test dataset and a training dataset. The training data set included the entire dataset except for the test data. The test dataset was a sample of the training dataset. Once the datasets were formatted, C4.5 could create rules for the decision tree and display the results. Once C4.5 read the datasets under the defaults the percentage of error was twenty percent for the test data. The defaults were set as weight = 2, and pruning confidence = 25%. By changing the weight to one and increasing the pruning confidence, percentages of error ranged from zero percent to twenty percent. One of the results found -15.9% of estimated percentage of error. After reviewing J. Ross Quinlan's book, the creator of C4.5, expressed that weight should be increased and pruning confidence should be decreased for noisy data. By taking Quinlan's suggestions, percentages of errors decreased from twenty percent to fifteen percent. There was one result that had a percentage of error of thirty percent. After analyzing the results that were based on Quinlan's suggestions and noticing the percentage of error was thirty percent, to receive optimum results using the default was the best option. Using the same strategy with the training data Quinlan's suggestions helped but did not produce optimum results as well as the default settings. After manipulating C4.5's weight and pruning confidence, the optimum results can be derived from the default with noisy and missing data. Once results from the default were displayed, it was apparent C4.5 was designed to outperform ID3.

Problem Description

This experiment was initially designed to assess C4.5 improvements from ID3. In order to test C4.5 three datasets were used. Each one is noisier than the next. In these datasets noisy means the data is difficult to create rules where the classes are clearly defined. The Diabetes dataset is considered difficult to create a nice decision tree because it contains missing values. Missing values in a dataset means that values in certain attributes are not existence. For the Diabetes dataset there was a training dataset and a test dataset. Both were formatted in comma delimited form. The other two datasets, Wine and Iris were going to be used as well in the comma delimited format. Once the datasets were formatted into the comma delimited format, the defaults of C4.5 could be manipulated to produce optimum results. After manipulating the weight and pruning confidence of C4.5

to produce optimum results, some results came to be negative percentages of error. Once results became skewed, the focus of the experiment switched to manipulating the weight and pruning confidence in other ways to discover new results. Instead of decreasing the weight and increasing the pruning confidence, Quinlan's suggestions were to increase the weight and decrease the pruning confidence. Taking Quinlan's suggestions, the percentages of errors became positive and small. Although Quinlan's suggestions decreased the percentage of error they still were not low enough. As a last chance the default was used to see if C4.5 could really handle these noisy datasets. By using the default settings, C4.5 produced even lower percentages of error.

Analysis Technique

While C4.5 is based on ID3, it was designed to outperform ID3 in many areas. The idea is to have C4.5 read two different datasets where one has a low percentage of noise and the second contains missing values and has a mid-high percentage of noise. By giving the algorithm datasets that can be hard to manipulate into appreciated results, we can compare the effectiveness of this algorithm to ID3.

First C4.5 needs to be able to read the dataset so the data has to be manipulated into a format where the algorithm can successfully interpret the data. After the algorithm reads the data, it performs a mathematical formula to decide where to start. The mathematical formula according to Wikipedia is:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

P is the probability of the dataset X belonging to class i. B represents log of base two. It will find the attribute with the highest probability in class S. It will split the attribute into two classes where the algorithm will split that information with another attribute. C4.5 will continue until most or all of the data instances are used. After C4.5 is finished, the results are used to determine if this algorithm can successfully split a dataset that has noise or missing values (Seidler, 2004).

C4.5 chooses an attribute with the highest gain or entropy to split and begin the decision tree. The algorithm continues to pick the second attribute with the next highest amount of gain to continue the tree.

C4.5 presents its results in a decision tree. When it splits an attribute into two classes, that attribute is at the top of the decision tree. The attribute is split into to two numerical values for an example it could split the attribute Class Intensity into less than or equal to 3.82 and greater than or equal to 3.81. The information that is under less than or equal to 3.82 will split with another attribute with two numerical values. Sometimes C4.5 will split an attribute into one numerical value depending on the dataset. Along with the decision tree it also gives the percentage of error of the decision tree where the

analyst can decide if they want to use that tree or manipulate the data to make the algorithm perform another tree with better results.

C4.5 is known to handle missing data and noisy data because it was designed to outperform ID3 (Dunham, 2006). According to the information on Wikipedia, C4.5 can deal with missing data because the values that are missing are not used in the entropy formula. Since the missing values are not used in the entropy formula they are not in the decision tree as well.

This project is designed to test C4.5 to see if it can outperform ID3.To test it against ID3, C4.5 will create two decision trees on a noisy dataset and a missing value dataset. Based on the results, it will be apparent if C4.5 was designed successfully or incorrectly.

Learning about C4.5 and its improvements on ID3, it seems as though C4.5 will be highly successful with the two datasets that were chosen. The hypothesis is that the decision trees will have a low percentage of error, but without any doubt C4.5 will create two successful decision trees. To test the hypothesis, the Wine and the Diabetes datasets were manipulated into a comma delimited format that was C4.5 friendly. Second, a training and a test set was created for both of the datasets. Third, the algorithm read the test sets for both of the datasets and created two decision trees. The percentage of error for the Wine dataset came to be 3.97% and for the Diabetes dataset it was 0%.

The wine dataset is a noisy dataset that does not have any missing values. The Wine dataset is difficult to construct rules where the classes are easily defined; since it does not have missing values, each value in each attribute is present. This dataset was perfect for testing C4.5 to see if it could handle noisy data correctly. In the Wine dataset there are 153 instances with 14 attributes. The data was obtained from three different farmers in Italy of the same region where the wines were analyzed chemically (Aleshunas). In the Diabetes dataset there are 768 instances with 9 attributes. The people that were examined were from the Pima Indian tribe (Aleshunas). The data was obtained by giving them physical examinations and personal interviews that centered around their medical history. The medical attributes ranged from whether they were pregnant to their body mass indexes. The Diabetes dataset is a noisy dataset with many missing values. Diabetes dataset is difficult to create clean classes, while some values in specific attributes are nonexistence. Since the Diabetes dataset contains missing values it creates a bigger obstacle than just having a noisy dataset. The Diabetes dataset was chosen because it contained two of the five aspects ID3 could not handle. These datasets were chosen to test C4.5 in the areas ID3 failed. These datasets had to be manipulated in order for C4.5 algorithm could read them. Both the Wine and the Diabetes dataset were edited to comma delimited format. They also had to be organized into training and test datasets in order to test the algorithm in a simple manner. The two test datasets are a few data instances that could represent the entire dataset for the Wine dataset and the Diabetes dataset. To test C4.5 completely, the entire dataset had to be manipulated also. The entire dataset could not include the test dataset. The data instances from the test dataset were deleted from the entire dataset of the Wine and the Diabetes data. By having a test and training (entire data subtract test data) dataset there were two ways to test C4.5.

Assumptions

The percentage of error is perceived as low considering the level of noise that is present in the Diabetes and the Wine datasets.

Results

By using the C4.5 defaults, the datasets Diabetes, Wine and Iris all produced optimum percentages of error. The Diabetes dataset produced a 7.5 percentage of error. The Wine dataset produced 0.7 percentage of error. The Iris dataset produced 2.7 percentage of error.

Issues

According to Quinlan large datasets (several hundred) when divided into training and test sets will result in large variations of error.

Appendices

There is not information related to this section.

References

Dunham, M. (2006). *Data Mining: Introductory and advanced topics*. India: Delhi, Dorling Kindersley Pvt. Ltd.

Quinlan, Ross J. (1993). *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufman Publishers, Inc.

C4.5. www. Mercury.webster.edu/aleshunas.

Seidler, T. (2004). *The C4.5 Project: An overview of the algorithm with results of experimentation.* Retrieved 12/7/09 from www.mercury.webster.edu/aleshunas

www.wikipedia.com Retrieved 11/27/09