Project 3: Decision Trees for Classification and Regression

Sarah Wilson   SWI1S117@JHU.EDU

303-921-7225

Engineering Professionals Computer Science

Johns Hopkins University

Baltimore, MD 21218, USA

1. Introduction

Decision trees are tree structures than can be built by machine learning algorithms by training on a data set. A decision tree is a hierarchical structure, consisting of nodes, which can have left or right nodes (children). The root node is the start of the tree and has left and right nodes that would go on to build sub-trees. A node that does not have a left or right child is considered a leaf. Decision trees are built or grown from the training data by deciding which features to build sub-trees on and knowing when to split the data to produce the next node, along with knowing when a leaf node has been encountered. Decision trees can be used in both classification and regression machine learning problems. Decision trees ask "questions" at each node and determine, based on feature observations from the data set, if the left or right node should be taken when traversing the tree. When the leaf node is encountered that will end with the answer for the classification or regression problem.

The problem statement presented in this paper is to understand and implement two major categories of decision tree algorithms. For classification problems the approach explored will be the Iterative Dichotomiser 3 (ID3) algorithm. For regression problems the approach explored will be the Classification and Regression Tree (CART) algorithm. Both ID3 and CART will first be built as standard uni-variate trees, meaning that they will be grown to completion with out intervention. The results of these trees will then be compared with trees that have been edited, for regression problems this will include the process of early stopping, for classification problems this will include the process of reduced error pruning. The uni-variate, early stopping and error pruning trees will be then compared in terms of Classification Error and Mean Square Error for classification and regression problems respectively. k-fold cross validation will be the experimental method used in this report.

The hypothesis of this report is that for the both the classification and regression tasks the trees that are error pruned or early stopped will perform better than trees that are grown uni-variate to completion. The hypothesis assumes that trees grown to full completion will exhibit over fitting to the data. Over-fitting is where in a learning system the model has been so tightly fit to the training data, that results on new un-trained data are highly inaccurate. The hypothesis assumes that in allowing the trees to grow to full completion of nodes that additional patterns may be learned in the training data that might not apply to new data feed into the algorithm. The results from experiments ran on the

provided data sets will be discussed and presented against the outlined hypothesis.

Section 1 has provided the introduction, problem statement and hypothesis in regards to two major decision tree algorithms that will be explored.. Section 2 will provide an in-depth explanation of the ID3 and CART algorithms, how these algorithms will be modified during the early stopping and error pruning processes and discuss each of the 6 data sets used. Section 3 will present the results obtained by variations of ID3 and CART as uni-variate, early stopping and error pruning implementations on the provided data sets. Section 4 will discuss the results that were obtained and compare them to the hypothesis that was outlined in the introduction. This report will conclude in Section 5 with a discussion of lessons learned and areas of possible future work.

2. Algorithms and Experimental Methods

The ID3 algorithm will be used on classification data sets. The ID3 algorithm is a classification algorithm that follows a greedy heurisitic top down approach by selecting the attribute/feature that provides the maximum information gain at each partition of the data set. The two parameters to understand in the ID3 algorithm are Entropy(I), Expected Entropy, Information Gain and Gain Ratio.
Entropy is xxxxx
Expected Entropy is xxxx
Information Gain is xxx
Gain Ratio

Consider a 2 class classification problem we can define the following equations for Entropy, Expected Entropy, Information Gain and Gain Ratio.

Entropy

$$I(p,n) = -\frac{p}{(p+n)} * log(\frac{p}{p+n}) - \frac{n}{p+n} * log(\frac{n}{p+n})$$

(1)

$I =$ Entropy of Partition

$\pi =$ Current Partition in the Overall Data Set

$p_\pi =$ Number of Class 1 Observations in the Current Partition

$n_\pi =$ Number of Class 2 Observations in the Current Partition

Expected Entropy

$$E_\pi(f_i) = \sum_{j=1}^{m_i}(\frac{p_\pi^j + n_\pi^j}{p_\pi + n_\pi} * I(p_\pi^j, n_\pi^j))$$

(2)

$E_\pi =$ Expected Entropy of Partition

$f_i =$ Feature i of the Data Set

$m_i =$ Size of the Domain of Feature i

$\pi =$ Current Partition in the Overall Data Set

$I =$ Entropy of Partition

$p_\pi^j =$ Number of Class 1 Observations in the Current Partition for the jth item in the Domain of Feature f_i

$n_\pi^j =$ Number of Class 2 Observations in the Current Partition for the jth item in the Domain of Feature f_i

Gain

$$G_\pi(f_i) = I(p_\pi, n_\pi) - E_\pi(f_i)$$

(3)

$G_\pi =$ Gain of Partition

$E_\pi =$ Expected Entropy of Partition

$I =$ Entropy of Partition

$p_\pi =$ Number of Class 1 Observations in the Partition

$n_\pi =$ Number of Class 2 Observations in the Partition

Information Value

$$IV_\pi(f_i) = -\sum_{j=1}^{m_i}(\frac{p_\pi^j + n_\pi^j}{p_\pi + n_\pi} * log\frac{p_\pi^j + n_\pi^j}{p_\pi + n_\pi})$$

(4)

$IV_\pi=$ Information Value of Partition

$m_i=$ Size of the Domain of Feature i

$p_\pi=$ Number of Class 1 Observations in the Partition

$n_\pi=$ Number of Class 2 Observations in the Partition

Gain Ratio

$$GR_\pi(f_i) = \frac{G_\pi(f_i)}{IV_\pi(f_i)} \tag{5}$$

$GR_\pi=$ Gain Ratio of Partition

$G_\pi=$ Gain of Partition

$IV_\pi=$ Information Value of Partition

These equations can be expanded to a **k** Class Classification problems as follows:
INSERT HERA

These equations will be used in the overall process of building the decision tree in the ID3 algorithm. The following steps outline the algorithm. The algorithm uses recursion to build the tree to completion.

The CART algorithm will be used on regression data sets.

Data Sets

The following data sets were used during the classification and regression tasks for this project.

Breast Cancer

Description:

Task: Classification

Predictor: Diagnosis (Malignant or Benign)

Link:

`https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29`

Car Evaluation

Description:

Task: Classification

Predictor: Car Evaluation (Unacceptable, Acceptable, Good, Very Good)

Link:

`https://archive.ics.uci.edu/ml/datasets/Car+Evaluation`

Congressional Vote

Description: 1984 United Stated Congressional Voting Records

Task: Classification

Predictor: Party (Republican / Democrat)

Link:

`https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records`

Albalone

Description: Physical measurements of Albalone

Task: Regression

Predictor: Rings (int)

Link:

`https://archive.ics.uci.edu/ml/datasets/Abalone`

Computer Hardware

Description: Relative CPU performance data.

Task: Regression

Predictor: PRP

Link:

`https://archive.ics.uci.edu/ml/datasets/Computer+Hardware`

Forest Fires

Description: Forest Fire burn area data

Task: Regression

Predictor: Area (float)

Link:

`https://archive.ics.uci.edu/ml/datasets/Forest+Fires`

3. Results

Tables 1-9 display the results from Classification Tasks: the Breast Cancer, Car Evaluation and Congressional Vote data sets while running Normal KNN, Edited KNN and Condensed KNN. These tables show the results from the train set and the test set during each fold of the k-fold validation process. The results of the tuning process were omitted from this report for brevity but can be found in the project submission, under Results Output. Each KNN variation was run using the optimal values as determined by the parameter tuning. Discussion of the tuning process will occur in the discussion section. The error used is classification error and can be described as the Number of Times Prediction was Wrong / Total Number of Comparisons. This error was averaged across the 5 folds to provide the Average Classification Error against each data set.

Tables 10-18 display the results from the Regression Tasks: the Albalone, Computer Hardware and Forest Fire data sets. These tables show the result from the train set and the test set during each fold of the k-fold validation process.The results of the tuning process were omitted from this report for brevity but can be found in the project submission, under Results Output. Each KNN variation was run using the optimal values as determined by the parameter tuning. Discussion of the tuning process will occur in the discussion section. The error used is Root Mean Square Error. Root Mean Squared Error was used, since it provides how far off a measurement was in the units of the measurement that was taken. The higher the Root Mean Squared Error, the worse the prediction was to the actual value. Low Root Mean Squared Error is the objective.

Table 1: Car Evaluation: ID3 - Experimental Results

Table 2: Breast Cancer: ID3 - Experimental Results

Table 3: Congressional Vote: ID3 - Experimental Results

Table 4: Albalone: CART - Experimental Results

Table 5: Computer Hardware: CART - Experimental Results

4. Discussion

5. Conclusion

6. References

1. Alpaydin, E. (2004). Introduction to machine learning (Oip). Mit Press.

Table 6: Forest Fires: CART - Experimental Results