Decision Trees Project Report

CS529 Machine Learning

Team Member:

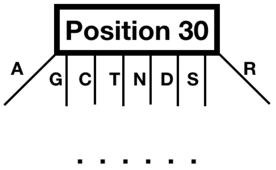
Chihyu Shen
Christina Xuan Yu

September 29, 2017

Part 1: Training

ID3 Implementation:

We adopted the ID3 algorithm to build the decision trees. Firstly, We chose a certain number of sample lines (for example 200) from training data, and built a 200 by 60 matrix, plus an extra column for the class. There are 3 classes total: EI, IE and N. For each attribute/position, it contains 8 possible values(A, G, C, T, N, D, S, R).



Secondly, we calculate the Information gain (using entropy or Gini Index) for every position (or column) from p1 to p60, and find the position with the max gain value, since we can get the most information from this position. We use this Max_GainValue_Position to be the first attribute/ node of the decision tree.

Thirdly, we repeat the above steps for the rest of the nodes.

Note, for every attribute generated, we adopt the Chi Square Test to check if this split is necessary.

Information Gain Using Entropy:

We first compute the total entropy for the whole set by the following formula:

$$H_{total} = -\sum_{i=1}^{3} P_i log P_i, \ where\ P_i\ is\ the\ probability\ of\ every\ possible\ class.$$

Then we compute the entropy for every attribute values, we use a slightly modified formula:

$$H_A = -\sum_{i=1}^{3} P_i log P_i, \ where\ P_i = \frac{|A \wedge Class_i|}{|A|}.$$

Then we use the same formula to compute entropy for the rest of the values. Lastly, we get the InfoGain value for each position with:

$$InfoGain(S,\ Position_x) = \ H_{total}\ - \sum_{i=1}^{8} \frac{|S_i|}{|S|} H(s_i)\},$$

$$where\ \frac{|S_i|}{|S|}\ is\ the\ ratio\ of\ each\ attribute\ value\ over\ the\ set.$$

Gini Index:

Gini index is another way to compute the information gain of the attribute. It is similar to the method above when we use entropy, we first compute Gini value for the whole set:

$$Gini_{total} = 1 - \sum_{i=1}^{3} P_i^2, \ where \ P_i = \frac{|Class_i|}{|S|}$$

Then we compute the gini value for every attribute values:

$$Gini_A = 1 - \sum_{i=1}^{3} P_i^2, \ where \ P_i = \frac{|A \wedge Class_i|}{|A|}.$$

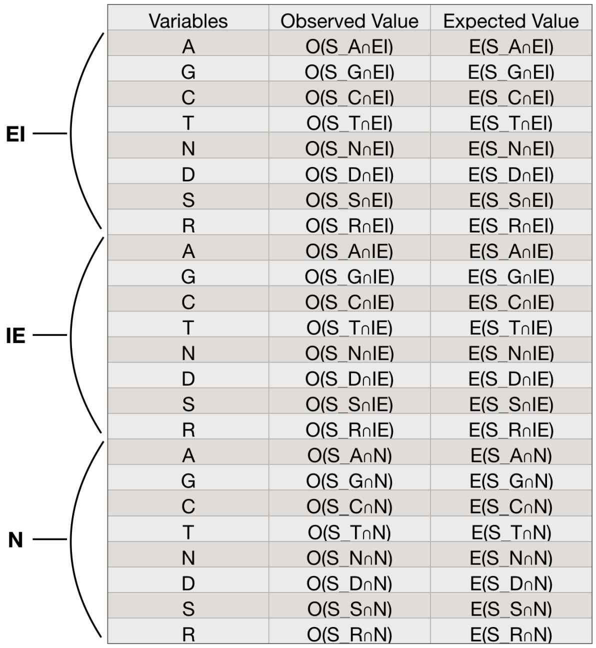
Then we use the same formula to compute Gini value for the rest of the attribute values. Lastly, we get the InfoGain for each position with:

$$InfoGain(S, \ Position_x) = Gini_{total} - \sum_{i=1}^{8} \frac{|S_i|}{|S|} Gini_{S_i},$$

$$where \ \frac{|S_i|}{|S|} \ is \ the \ ratio \ of \ each \ attribute \ value \ over \ the \ set.$$

Chi Square Test With Different Confidence Levels:

To make sure the attribute growing as a node on the tree is actually providing us with additional information, not just having a large percentage by chance, we implement Chi square test to every attribute generated.
For every attribute (Position x), we count the number of times each attribute value occurs with respect to each class as the observed real count.



	Variables	Observed Value	Expected Value
EI	A	O(S_A∩EI)	E(S_A∩EI)
	G	O(S_G∩EI)	E(S_G∩EI)
	C	O(S_C∩EI)	E(S_C∩EI)
	T	O(S_T∩EI)	E(S_T∩EI)
	N	O(S_N∩EI)	E(S_N∩EI)
	D	O(S_D∩EI)	E(S_D∩EI)
	S	O(S_S∩EI)	E(S_S∩EI)
	R	O(S_R∩EI)	E(S_R∩EI)
IE	A	O(S_A∩IE)	E(S_A∩IE)
	G	O(S_G∩IE)	E(S_G∩IE)
	C	O(S_C∩IE)	E(S_C∩IE)
	T	O(S_T∩IE)	E(S_T∩IE)
	N	O(S_N∩IE)	E(S_N∩IE)
	D	O(S_D∩IE)	E(S_D∩IE)
	S	O(S_S∩IE)	E(S_S∩IE)
	R	O(S_R∩IE)	E(S_R∩IE)
N	A	O(S_A∩N)	E(S_A∩N)
	G	O(S_G∩N)	E(S_G∩N)
	C	O(S_C∩N)	E(S_C∩N)
	T	O(S_T∩N)	E(S_T∩N)
	N	O(S_N∩N)	E(S_N∩N)
	D	O(S_D∩N)	E(S_D∩N)
	S	O(S_S∩N)	E(S_S∩N)
	R	O(S_R∩N)	E(S_R∩N)

We compute the expected values for each attribute value with respect to each class as follows:

$$E(S_{X_i \wedge Y_j}) = |X_i|\frac{|Y_i|}{|S|}$$

Then, with all the numbers on the table ready, we compute the critical value c, it follows the formula:

$$C = \sum_{j=1}^{3} \frac{(Observed_{X_i}{}^{j} \; - Expected_{X_i}{}^{j})^2}{Expected_{X_i}{}^{j}}$$

Since we have 3 classes and 8 attribute values (also attempted with 4 values), we get the degree of freedom DOF = (NumClass - 1)*(NumAttrValues - 1) = 14.

With the confidence level of 95% (also attempted 99% and 0%) for example, we can find the X value (23.685) on the Chi Square Test Table according to the DOF being14.

Then we compare our critical value with the X above to make our decision:

1. If the critical value c ≤ X, then we determine the split is by chance, we stop this branch from growing. And we use the label instead of an attribute node.
2. If the critical value c > X, then we determine that the split is not by chance, we let this branch keep growing.

Part 2: Testing

10-Fold Cross Validation and Various confidence levels:

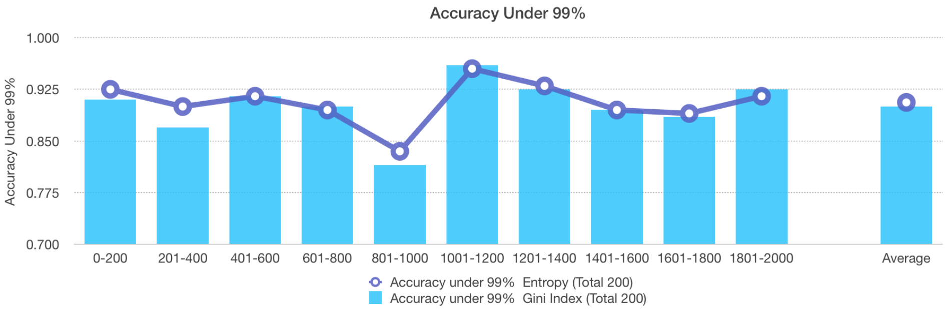
We have compared the accuracies with 99%, 95% and 0% with 10-fold cross validation. We cut 2000 training lines into 10 parts. Each single part was used as validation while the other 9 parts are used as training data.

x% means that we set the confidence interval so that there is a x% chance that it contains the true population mean. Among all the testings, 95% confidence level provided us with the highest accuracy from testing. And 99% is a little bit lower but close to it. With 0% confidence level, we basically implemented the tree without the Chi square test. Besides the accuracy was significantly lower than the other two—average accuracy is around 0.8, the decision trees generated with 0% were overly large. As we know, the ideal decision tree should have the maximum depth no more than 60—with 60 DNA sequence positions only per sample line. If a tree has the depth over 60, there are unnecessary nodes generated, which may be misleading and lower the accuracy of classification.

The test results are provided on the next page.

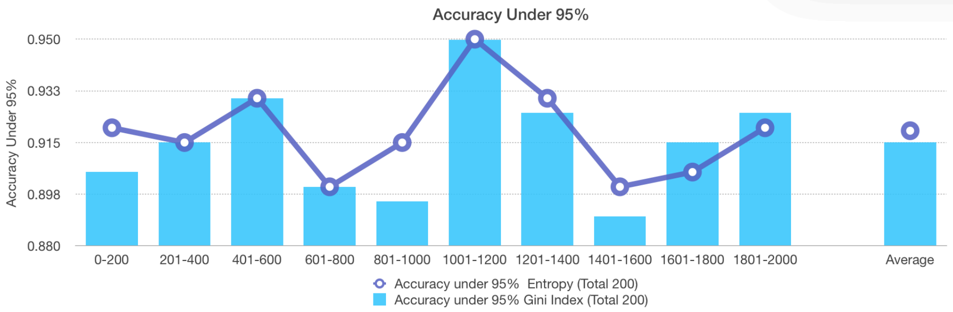
10-Fold Cross Validation Under Different Parameters

	ACCURACY UNDER 99% ENTROPY (TOTAL 200)	ACCURACY UNDER 99% GINI INDEX (TOTAL 200)
0-200	0.925	0.910
201-400	0.900	0.870
401-600	0.915	0.915
601-800	0.895	0.900
801-1000	0.835	0.815
1001-1200	0.955	0.960
1201-1400	0.930	0.925
1401-1600	0.895	0.895
1601-1800	0.890	0.885
1801-2000	0.915	0.925
Average	0.906	0.900



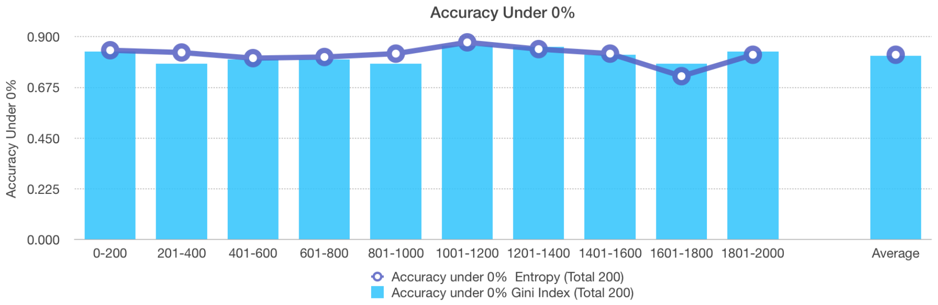
10-Fold Cross Validation Under Different Parameters

	ACCURACY UNDER 95% ENTROPY (TOTAL 200)	ACCURACY UNDER 95% GINI INDEX (TOTAL 200)
0-200	0.920	0.905
201-400	0.915	0.915
401-600	0.930	0.930
601-800	0.900	0.900
801-1000	0.915	0.895
1001-1200	0.950	0.950
1201-1400	0.930	0.925
1401-1600	0.900	0.890
1601-1800	0.905	0.915
1801-2000	0.920	0.925
Average	0.919	0.915



10-Fold Cross Validation Under Different Parameters

	ACCURACY UNDER 0% ENTROPY (TOTAL 200)	ACCURACY UNDER 0% GINI INDEX (TOTAL 200)
0-200	0.840	0.835
201-400	0.830	0.780
401-600	0.805	0.800
601-800	0.810	0.800
801-1000	0.825	0.780
1001-1200	0.875	0.860
1201-1400	0.845	0.855
1401-1600	0.825	0.820
1601-1800	0.725	0.780
1801-2000	0.820	0.835
Average	0.820	0.815



Information Gain with Entropy VS Gini Index Test result Comparison:

From those comparison tests, we can also see that the accuracies generated by the tree implementing infoGain with entropy and the tree implementing Gini Index are very close. With CL 95%, between line 1601 and 2000, we got a higher accuracy from Gini index. And in some other sections like line 0-200 and 801-1000, infoGain with entropy got a more accurate result. Generally, they both work well with finding the attribute with the most information provided. The entropy is a way to measure impurity. And the Gini index is a criterion to minimize the probability of misclassification. They both help to decide which attribute to select at each step in building the tree.

Other Things We Tried:

1. We tried to set the attribute values to 4(A, G, C, T) instead of 8 (A, G, C, T, N, D, S, R). It worked well for some sample lines. However after multiple tests, we found using all 8 attribute values are more generalized, especially in the Chi-square test. We get a higher accuracy up to 91% using 8 attribute values on the final test file, while the accuracy rate is between 70%-89% using 4 values.
2. We trained and tested the data with the whole set, and got a result from it— 1884 were correct out of 2000. Then we analyzed the result, took off the samples that we received an incorrect classification. So we know the extreme cases which could be misleading were omitted. Using the remaining samples, we build another tree. Finally we used this tree on the final test file, and the accuracy is the highest among all of our tests—91.428%
3. We tried to use the randomized sample lines to train the data, and 4-fold cross validation. And their accuracies range between 70%-90%.

Part 3: Summary

Implementing ID3 decision tree, information gain with entropy and Gini index are both good ways to generate the nodes with the most information provided. Chi Square Test helps us to split the tree when necessary. Applying 99% and 95% of confidence level results in close accuracies, while 95% provides a slightly better result. 0% was not a good confidence level in our tests, it produced larger trees with lower levels of accuracy.

Our best score for the final test is **91.596%** achieved by using the tree generated under the parameters of 10-fold cross validation, lines 1000-1200 as testing data, the rest of the 1800 lines as training data, information gain with entropy and a 95% confidence level.