Kishore Kumar Vaddineni
UNM ID: 101671250

CS-529: Project-1

Aim

Implement ID3 algorithm with chi square stopping criterion.

Abstract

In this project we will implement a decision tree algorithm and apply it to molecular biology, a Promoter is a region of DNA that facilitates the transcription of a particular gene. The task for you is to develop a decision tree algorithm, learn from data, and predict for unseen DNA sequences whether they are promoters or non-promoters.

ID3 is decision tree learning algorithm, which formulates a decision tree from a given training set. This algorithm determines most promising attributes which will help classify unseen data records and generates a tree. Chi square statistical value is used to stop growing tree based on hypothesis of independency of attribute on record. Early stopping increases over-fitting of records.

Description of implemented code

Following storage representations are used in the Code:-

- Output Label
  - '+' as PROMOTER
  - '-' as NON_PROMOTER
- Attributes are represented by Attribute Index (Attribute_column#) for each record, index starting with 0. For e.g. In 'cgta', AI of 'c' is 1, AI of 't' is 3 and so on and in the output it is showed as Attribute_1 for 'c' and Attribute_3 for 't'.
- Record is represented as an object, which is stored in List and record's assigned Output Label.
- Attribute is also stored in a List and is also having the labels assigned.
- Chi square value is represented with the variable chisquarevalue and a default value of '0' was assigned to it.

Algorithm of code

1. Load 'training.txt' and 'validation.txt' into 'validationfile' and 'datafile' ArrayList of Record Class.
2. After loading the data assign the labels to the attributes.
3. Then divide the attributes from the promoters/nonpromoters column.
4. Using this dimension labels and Examples, these values are passed to ID3 Algorithm Method.
5. Calculate the Information Gain for each node.

6. Calculate the Misclassification error for a node.
7. The above methods are used to generate the tree.
8. This step will be repeated until the complete decision tree is generated.
9. After the generation of complete tree, purning will be done in reverse order.
10. After the tree has been generated the validation data is used to classify the DNA samples.
11. The above steps will be repeated to train the classifier using the training data provided and find the accuracy on the validation data by varying the confidence level with various values such as 99%, 95%, and 0%.

<u>Results obtained under various settings</u>

Accuracy obtained using under various cases

S.No	Degree of Freedom	Chi-square Table Value	Confidence Values	Accuracy
1	3	0	0%	85.714
2	3	7.815	95%	80.0
3	3	11.345	99%	82.85

<u>Question</u>: Why does the accuracies change for each confidence value? Let’s see the reason behind it.

At 0% confidence level: The chi-square test is not applied to the nodes, there is no pruning at all, and hence the tree will grow completely. At 95% confidence level: The chi-square table value is 7.815. After computing the chi-square value for each node, since the threshold is small, and thus there is a possibility for large number of nodes to be pruned. Hence the accuracy will be low since the node which is pruned might be valuable. At 99% confidence level: The chi-square table value is 11.345, this threshold is large comparatively to the above 95% threshold, and thus there is a possibility for less number of nodes to be pruned. Hence the accuracy will be higher than the above specified confidence (95%) but definitely less than 0% confidence.