Indian Institute of Technology Ropar

CSL603

Decision Tree Learning

Submitted To:
Narayanan C Krishnan
Computer Science
Department

Submitted By :
Siddharth Nahar
2016CSB1043

Contents

1 Decision Tree Model and Analysis

This is Report for Decision Tree Learning through Algorithm ID3. In this Assignment we have to Classify Movie Reviews using Word Sentiments as Features taken from http://ai.stanford.edu/ amaas/data/sentiment/index.html.

I have Sampled 1000 random subset from Train Data Set and Test Data Set and Selected top 2500 positive sentiment Attributes and bottom 2500 negative sentiments Attributes and created Feature vector in TrainSet.txt and TestSet.txt and selected Features in selected-features-indices.txt . All the Results for Analysis and Accuracy has been Discussed below.

2 Basic Analysis of Decision Tree Created

This Analysis is conducted for Binary Splitting of Nodes and Selection of Attribute on basis of Information Gain.

Formulas Used for calculation of Entropy and Information Gain :

Let Entropy = E and p_i = Probablity of i^{th} class.

$$E = \sum_{i=1}^{c} p_i * log(p_i)$$

Let Information Gain w.r.t to A = Gain(S,A)

$$Gain(S, A) = Entropy(S, A) - \sum_{i=1}^{v} \frac{|S_i|}{|S|} * Entropy(S_i, A)$$

$$WeightedEntropy(S, A) = \sum_{i=1}^{v} \frac{|S_i|}{|S|} * Entropy(S_i, A)$$

$$AttributeSelection = \min_{\forall A_i \in A} WeightedEntropy(S, A_i)$$

2.1 Accuracy and Analysis for Decision Trees

Early stopping Criteria Used are :

1. Weighted Entropy :- 6th Column Corresponds to stopping Criteria on Entropy to Train model.
2. Height of Tree :- 7th Column corresponds to Height as stopping criteria.
2.a Max Height = 180 for 5000 Attributes
2.b Max Height = 100 for 3000 Attributes
2.c Max Height = 40 for 2000 Attributes

	Sample	Train Acc	Test Acc	Height	Weighted Entropy	Stopping Height
Attributes = 5000	1	89.1	73.8	216	$\leq 0.1 - 75.0$	73.8
	2	89.9	71.2	208	$\leq 0.2 - 71.8$	72.0
	3	87.1	67.0	221	$\leq 0.3 - 71.3$	67.0
	4	89.6	68.7	216	$\leq 0.4 - 72.4$	69.0
Attributes = 3000	1	82.9	73.8	148	$\leq 0.1 - 73.9$	73.8
	2	84.0	73.6	157	$\leq 0.2 - 74.1$	73.9
	3	83.4	69.3	153	$\leq 0.3 - 69.3$	69.3
	4	83.3	70.3	158	$\leq 0.4 - 71.7$	70.3
Attributes = 2000	1	75.4	69.3	104	$\leq 0.1 - 69.3$	68.4
	2	74.2	70.6	103	$\leq 0.2 - 71.1$	69.4
	3	76.6	69.4	116	$\leq 0.3 - 69.4$	68.8
	4	75.9	70.0	122	$\leq 0.4 - 70.0$	69.3

5000 Attributes		3000 Attributes		2000 Attributes	
words	Frequency	words	Frequency	words	Frequency
poor	10	avoid	8	pathetic	5
avoid	6	terrible	5	garbage	5
dull	6	poorly	5	crap	5
horrible	6	laughable	5	terrible	5
terrible	5	insult	5	costs	4
poorly	5	mess	5	waste	3
dumb	5	trash	5	unfunny	3
disgusting	5	stupid	4	insulting	3
painful	5	horrible	4	appalling	3
wasted	4	lame	4	sucks	3

2.2 Accuracy and Analysis for Decision Trees after Addition of Noise

In this step of Analysis Various percent of noise are added.Train Data Labels are randomly changed.Various Observation are Observed as follows :

- Height of Tree remains or small increase is observed. As Decision Tree is robust for noise. But due to noise it may overfit the data and reduce accuracy on test set.
- Due to bias partition on attribute value less data may be misclassified.So height changes very small.

			Before Adding Noise			After Adding Noise		
		Sample	Training Error	Test Error	Height	Training Error	Test Error	Height
Noise = 0.5%	Attributes = 5000	1	88.7	71.3	204	88.3	71.1	204
		2	88.7	67.7	189	88.4	67.6	189
	Attributes = 3000	1	83.5	73.2	159	83.1	73.3	164
		2	82.6	71.7	151	82.6	71.7	151
	Attributes = 2000	1	76.5	70.1	135	76.3	69.9	135
		2	76.4	67.9	129	76.4	67.9	129
Noise = 1%	Attributes = 5000	1	89.4	67.9	225	89.0	68.3	222
		2	88.7	66.8	196	87.7	65.6	193
	Attributes = 3000	1	81.9	72.0	149	81.6	71.6	153
		2	83.0	68.4	170	82.6	68.4	170
	Attributes = 2000	1	75.2	68.5	104	74.9	68.5	104
		2	78.6	66.7	138	78.5	66.9	138
Noise = 10%	Attributes = 5000	1	89.8	71.1	181	84.3	70.3	200
		2	90.3	68.2	194	83.6	67.3	212
	Attributes = 3000	1	84.0	70.7	153	79.9	68.5	157
		2	83.7	72.0	154	80.3	71.2	150
	Attributes = 2000	1	77.2	70.4	125	75.1	70.3	125
		2	75.9	70.8	121	73.7	70.1	125

3 Pruning of Trees

In this step of Analysis we Use Post-Pruning Strategy to learn the Tree.In Post-Pruning we learn decision tree without Early Stopping and Prune Tree to get Optimal Solution of Tree. For Different number of Attributes Observations are :

- With Pruning Height of Tree Decreases considerably.
- Accuracy on Test Data Increases and Training Error is constant or increases.
- Observation has been Absorbed on various samples and on different attributes.

	Sample	Without Pruning			With Pruning		
		Train Acc	Test Acc	Height	Train Acc	Test Acc	Height
Attributes = 5000	1	88.5	68.8	209	88.5	72.0	89
	2	88.7	67.2	207	88.7	68.5	95
	3	89.6	71.2	197	89.6	71.8	94
	4	89.8	69.5	216	89.8	73.8	87
Attributes = 3000	1	85.0	72.1	141	85.0	72.1	71
	2	82.7	73.5	160	82.7	75.0	76
	3	83.4	71.5	190	83.4	72.0	81
	4	83.1	69.5	159	83.1	72.1	80
Attributes = 2000	1	77.0	68.9	128	77.0	69.3	115
	2	75.9	69.0	114	75.9	69.3	57
	3	77.0	71.8	112	77.0	71.8	63
	4	76.3	66.7	111	76.3	66.8	67

4 Feature Bagging

In this step of Analysis we Randomly select some number of Attributes and Train model of Decision Tree as Number of Trees as Parameter.Various Observation are as follows :

- In feature Bagging we are selecting 2000 attributes Randomly So Decision Tree may not get well train.
- As Tree number increases Accuracy increases as more Trees are trained and classified well.
- For small number of forest size , accuracy is less than original accuracy.

Number of Attributes Selected = 2000
Before Random Forest Training Accuracy = 89.4
Before Random Forest Testing Accuracy = 68.6

No. of Trees	Training Accuracy	Testing Accuracy
5	79.8	68.3
10	78.3	67.6
20	81.2	69.0
30	82.1	69.2
50	84.0	71.0