Autonomus University of Queretaro

Faculty of Engineering
Master Science in Artificial Intelligence

Implementation of ID3 and CART Algorithms.

Student:
Iván Alejandro García Amaya

Instructor:
Phd. Marco Antonio Aceves Fernández.

Course:
Machine Learning.

Contents

List of Figures

List of Tables

1 Introduction.

In the following work an introduction to two tree search algorithms is made, the first one is ID3 and the second CART, for both algorithms the same database is used, only the authors handle a small difference in one of the attributes and It calls it differently, for example the first one calls the first attribute Weather and the second author calls it Outlook, within the instances the first names a Cloudy and the second Overcast, outside of these two small differences, it is understood that the dataset It does not matter.

In the figure 1, the conventional structure of a search tree is observed.

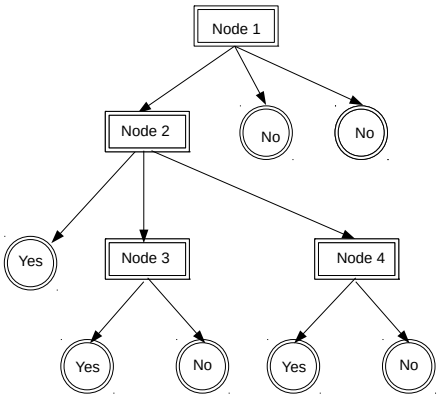

Figure 1: Decision tree.

The figure shows the dataset shown in the work Implementation of ID3 Algorithm by Rupali Bhardwaj that collected the state of the weather for 15 days, he mentions that the problem is based on finding a good day to play Cricket.

Day	Weather	temp	Humidity	Wind	Play
1	Sunny	85	85	week	No
2	Sunny	80	90	Strong	No
3	Cloudy	83	78	Week	Yes
4	Rainy	70	96	Week	Yes
5	Rainy	68	80	Week	Yes
6	Rainy	65	70	Strong	No
7	Cloudy	64	65	Strong	Yes
8	Sunny	72	95	Week	No
9	Sunny	69	70	Week	Yes
10	Rainy	75	80	Week	Yes
11	Sunny	75	70	Strong	Yes
12	Cloudy	72	90	Strong	Yes
13	Cloudy	81	75	Week	Yes
14	Rainy	71	85	Strong	No

Table 1: Dataset.

It mentions that discrete values are needed like the ones shown in the table 2

Attribute	Posible values A	Posibles values B	Posibles values C
Weather	Sunny	Cloudy	Rainy
Temperature	Hot	Medium	Cold
Humidity	Hgh	Normal	
Wind	Strong	Week	
Class	Play	No play	
Decision	N(negative)	P(positive)	

Table 2: Acceptable values.

You need to convert the Temperature and Humidity attributes to discrete values such as hot, medium, cold, high, and normal.

Temperature	Hot (H) 80 to 85	Medium (M) 70 to 75	Cold (C) 64 to 69
Humidity	High (H) 81 to 96	Normal (N) 65 to 80	
Class	Yes (Y) Play	No (N) no play	

Table 3: Discrete values .

In the table 4 it is observed with the discrete values added.

Day	Weather	temp	Humidity	Wind	Play
1	Sunny	85 Hot	85 High	week	No
2	Sunny	80 Hot	90 High	Strong	No
3	Cloudy	83 Hot	78 High	Week	Yes
4	Rainy	70 Medium	96 High	Week	Yes
5	Rainy	68 Cold	80 Normal	Week	Yes
6	Rainy	65 Cold	70 Normal	Strong	No
7	Cloudy	64 Cold	65 Normal	Strong	Yes
8	Sunny	72 Medium	95 High	Week	No
9	Sunny	69 Cold	70 Normal	Week	Yes
10	Rainy	75 Medium	80 Normal	Week	Yes
11	Sunny	75 Medium	70 Normal	Strong	Yes
12	Cloudy	72 Medium	90 High	Strong	Yes
13	Cloudy	81 Hot	75 Normal	Week	Yes
14	Rainy	71 Medium	85 High	Strong	No

Table 4: Dataset complete.

2 Methodology

2.1 ID3 Algorithm.

2.1.1 Entropy.

The level of uncertainty, termed entropy is:

$$-\sum_i P_i log_2 P_i \tag{1}$$

The frequency of event can be used as a probability estimate, suppose if exist 8 positive events and 2 negatives events, the probability for this node is given by:

$$\frac{8}{10} = 0.8 \tag{2}$$

2.1.2 Calculation of entropy.

The first instance is Weather where 14 examples are 9 Yes and 5 No, then

$$Entropy(S) = -\frac{9}{14} log_2(\frac{9}{14}) - (\frac{5}{14}) log_2(\frac{5}{14}) = 0.940 \tag{3}$$

Then the entropy is given by:

$$Entropy(S_{sunny}) = -\frac{2}{5} log_2(\frac{2}{5}) - (\frac{3}{5}) log_2(\frac{3}{5}) = 0.97095 \tag{4}$$

$$Entropy(S_{cloudy}) = -\frac{4}{4} log_2(\frac{4}{4}) - (\frac{0}{4}) log_2(\frac{0}{4}) = 0 \tag{5}$$

$$Entropy(S_{rainy}) = -\frac{3}{5} log_2(\frac{3}{5}) - (\frac{2}{5}) log_2(\frac{2}{5}) = 0.970950 \tag{6}$$

$$Gain(S_{Weather}) = Entropy(S) - \frac{5}{14} Entropy(S_{sunny}) - (\frac{4}{14}) Entropy(S_{cloudy}) - (\frac{5}{14}) Entropy(S_{rainy}) \tag{7}$$

$$Gain(S_{Weather}) = 0.940 - (\frac{5}{14}) 0.97095059 - (\frac{4}{14}) 0 - (\frac{5}{14}) 0.97095059 = 0.246 \tag{8}$$

The next step is to calculate the entropy of the attribute of Temperature, we know from table 4 that temperature = hot is of occurrences 4, medium is occurrence 6 and cold occurence of 6 .

$$Entropy(S_{hot}) = -\frac{2}{4}log_2(\frac{2}{4}) - (\frac{2}{4})log_2(\frac{2}{4}) = -0.9999 \tag{9}$$

$$Entropy(S_{medium}) = -\frac{4}{6}log_2(\frac{4}{6}) - (\frac{2}{6})log_2(\frac{2}{6}) = -0.91829583 \tag{10}$$

$$Gain(S_{cold}) = -\frac{3}{4}log_2(\frac{3}{4}) - (\frac{1}{4})log_2(\frac{1}{4}) = -0.81127812 \tag{11}$$

$$Gain(S_{Temp}) = Entropy(S) - (\frac{4}{14})Entropy(S_{hot}) - (\frac{6}{14})Entropy(S_{medium}) - (\frac{4}{14})Entropy(S_{cold}) \tag{12}$$

$$Gain(S_{Temp}) = 0.940 – (4/14)0.99999 – (6/14)0.91829583 – (4/14)0.81127812 = 0.0289366072 \tag{13}$$

Algorithm 1: Entropy .

input : dataset
output: *entropy*

```
n ← Number of attributes;
for i → n do
    'sunny','cloudy','rainy' ← uniqueValues(i, data1);
    a ← 'sunny';'cludy';'rainy';
    for j → a do
        b ← aj, 1);
        NumYes,NumNo ← countYesNo(b, data1);;
        entropy ← getEntropy(NumYes, NumNo);
    end
    entropyVec ← entropy;
end
```

The table 5 shows the results applying the procedure passed in the different attributes.

Entropy (S)	0.940
Gain(S. weather)	0.246
Gain(S. Temp)	0.0289366072
Gain(S. Humidity)	0.1515496
Gain(S. Wind)	0.048

Table 5: Entropy.

The one with the highest result is Weather, so it has 3 possible values: Sunny, Cloudy, Rain, for example $S_{sunny} = \{D1, D2, D8, D9, D11\}$

A subset is generated like the one shown in the table 6, where we will apply the same procedure as described above.

Day	Weather	Temp	Humidity	Wind	Play
1	Sunny	Hot	High	Week	No
2	Sunny	Hot	High	Strong	No
3	Sunny	Medium	High	Week	No
4	Sunny	Cold	Normal	Week	Yes
5	Sunny	Medium	Normal	Strong	Yes

Table 6: Subset A.

After applying the same procedure but with the use of the subset, results like those shown in the table 7are obtained.

Entropy(S_{sunny})	0.970
Gain(S_{sunny}. Temp)	0.057
Gain(S_{sunny}. Humidity)	0.971
Gain(S_{sunny}. Wind)	0.02

Table 7: Results subset A.

The next step is to get the results for $S_{Cloudy} = \{D3D7D12D13\}$.

Day	Weather	Temp	Humidity	Wind	Play
3	Cloudy	Hot	High	Week	Yes
7	Cloudy	Cold	Normal	Strong	Yes
12	Cloudy	Medium	High	Strong	Yes
13	Cloudy	Hot	Normal	Week	Yes

Table 8: Subset B.

In the table 9, it is observed that the result for these different values is zero, this means that for all possible cases it is "Yes"

Entropy(S_{cloudy})	0
Gain(S_{cloudy} Temp)	0
Gain(S_{cloudy} Humidity)	0
Gain(S_{cloudy} Wind)	0

Table 9: Results subset B.

To search the results with $Rainy = \{D4D5D6D10D14\}$

Day	Weather	Temp	Humidity	Wind	Play
4	Rainy	Medium	High	Week	Yes
5	Rainy	Cold	Normal	Week	Yes
6	Rainy	Cold	Normal	Strong	No
10	Rainy	Medium	Normal	Week	Yes
14	Rainy	Medium	High	Strong	No

Table 10: Subset C.

The table 11 shows that the one with the highest entropy is Wind

Entropy (S_{Rainy})	0.9710
Gain(S_{Rainy}. Temp)	0.0200
Gain(S_{Rainy}. Humidity)	0.020
Gain(S_{Rainy}. Wind)	0.9710

Table 11: Results subset C.

2.1.3 CART

2.1.4 Gini Index.

Gini Index is a metric classification tasks in CART. It stores sum of squared probabilities of each class. We can formulate it as illustrated below.

$$Gini = 1 - \sum(Pi)^2 \tag{14}$$

2.1.5 Outlook.

Outlook is a nominal feature. It can be Sunny, Claudy or Rainy. This will be summarize the final decision for features

Outlook	Yes	No	Number of instances
Sunny	2	3	5
Overcast (Cloudy)	4	0	4
Rainy	3	2	5

Table 12: Features

$$Gini(Outlook = Sunny) = 1–(\frac{2}{5})^2–(\frac{3}{5})^2 = 1–0.16–0.36 = 0.48 \tag{15}$$

$$Gini(Outlook = Overcast) = 1–(\frac{4}{4})^2–(\frac{0}{4})^2 = 0 \tag{16}$$

$$Gini(Outlook = Rain) = 1–(\frac{3}{5})^2–(\frac{2}{5})^2 = 1–0.36–0.16 = 0.48 \tag{17}$$

Then, we will calculate weighted sum of gini indexes for outlook feature.

$$Gini(Outlook) = (\frac{5}{14})0.48 + (\frac{4}{14})0 + (\frac{5}{14})0.48 = 0.171 + 0 + 0.171 = 0.342 \tag{18}$$

2.1.6 Temperature.

Similarly, temperature is a nominal feature and it could have 3 different values: Cool, Hot and Mild, summarize decisions for temperature feature.

Temperature	Yes	No	Number of instances
Hot	2	2	4
Cool	3	1	4
Mild	4	2	6

Table 13: Temperature Features

$$Gini(Temp = Hot) = 1–(\frac{2}{4})^2–(\frac{2}{4})^2 = 0.5 \tag{19}$$

$$Gini(Temp = Cool) = 1–(\frac{3}{4})^2–(\frac{1}{4})^2 = 1–0.5625–0.0625 = 0.375 \tag{20}$$

$$Gini(Temp = Mild) = 1–(\frac{4}{6})^2–(\frac{2}{6})^2 = 1–0.444–0.111 = 0.445 \tag{21}$$

We will calculate weighted sum of gini index for temperature feature

$$Gini(Temp) = (\frac{4}{14})0.5+(\frac{4}{14})0.375+(\frac{6}{14})0.445 = 0.142+0.107+0.190 = 0.439 \tag{22}$$

2.1.7 Time to decide.

We've calculated gini index values for each feature. The winner will be outlook feature because its cost is the lowest.

Feature	Gini Index
Outlook	0.342
Temperature	0.439
Humidity	0.367
Wind	0.428

Table 14: Features Gini Index

Algorithm 2: Entropy .

input : dataset
output: *entropy*

$n \leftarrow Number\ of\ attributes$;
for $i \rightarrow n$ **do**
  $'sunny','cloudy','rainy' \leftarrow uniqueValues(i, data1)$;
  $a \leftarrow 'sunny';'cludy';'rainy'$;
  for $j \rightarrow a$ **do**
    $b \leftarrow aj, 1)$;
    $NumYes,NumNo \leftarrow countYesNo(b, data1)$;;
    $Gini \leftarrow getGiniIndex(NumYes, NumNo)$;
  end
  $GiniVec \leftarrow Gini$;
end

3 Results and discussion.

The following image shows the result obtained using the ID3 algorithm

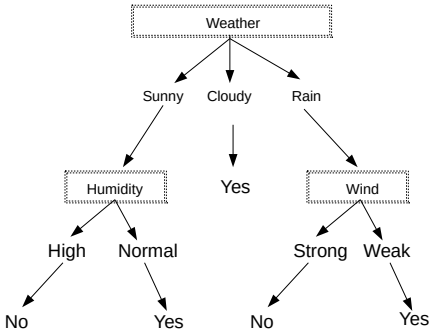

Figure 2: Weather tree.

In this decision tree we observe the one generated by the CART algorithm.

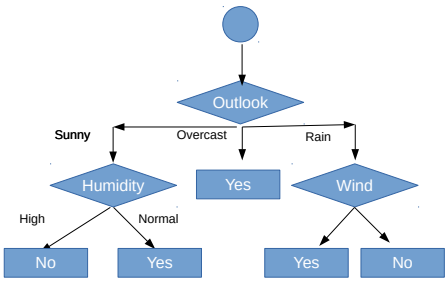

Figure 3: Weather tree cart.

The final results:

- If the weather is sunny and humidity is normal then play cricket but if humidity is high do not play cricket.
- If the weather is Cloudy then play cricket.
- If the weather is rainy and the wind is weak then play cricket but if wind is strong do not play cricket

4 Conclusion and Future Work

We can observe the two different types of algorithms that are presented in both ID3 and cart, in the methodology section, it is observed that both algorithms present great similarity, but the difference is that one in this case ID3 is based on obtaining the entropy for making decisions, while cart does it through the evaluation of the Gini variable, in the results section we observe the two different types of decision trees, these show similar or equal results, only that in these two they are used some different names.

Only ID3 decides the one with the lowest entropy while CART decides for the one that obtains a higher Gini value, on the other hand in complexity or execution time that are also very similar.

5 References.

Rupali Bhardwaj, Sonia Vatta. (2013). Implementation of ID3 Algorithm. International Journal of Advanced Research in Computer Science and Software Engineering, 3, 4-9.

Sefik Ilkin Serengil. (2018). A Step by Step CART Decision Tree Example. 2020, de Sefiks Sitio web: https://sefiks.com/2018/08/27/a-step-by-step-cart-decision-tree-example/