Comparison of SPRINT and PUBLIC

Xiang Wan
Qiang Ye

Dec 11, 2000

Content

- Introduction
- SPRINT Algorithm
- PUBLIC Algorithm
- Training Dataset
- Experimental Results
- Future Work

Advantage of Decision Tree

- Classification Techniques
  -Decision trees
  -Bayesian classification
  -Neural networks …
- Advantage of Decision Tree
  -Easily comprehended by humans.
  -Efficient and thus suitable for large training sets.
  -Do not require additional information

Building a Decision Tree

- Building Phase
  The tree is built by recursively partitioning the dataset until each partition is “pure”.
- Pruning Phase
  After fully grown, the tree is pruned to remove noises that may be particular only to the training datasets .

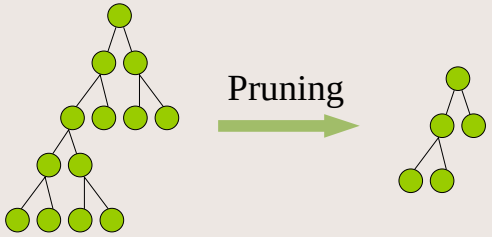
SPRINT and PUBLIC

- Using large training datasets, we can improve the accuracy of classification model
- Traditional decision tree algorithms(ID3 and C4.5) are established for small datasets

- SPRINT : **Building Phase**+**Pruning Phase**
- PUBLIC : **Integrating** the second "**Pruning Phase**" with the initial "**Building Phase**"

Why PUBLIC

- Building Phase takes more than **99%** of the total time needed to build a tree
- Pruning Phase prunes large portions(sometimes 90%) of the original tree



Initial Tree Final Tree

Attribute Lists

- One **attribute list** is for one attribute in the datasets
- Each record in the attribute list consists of an **attribute value**, a **class label** and the **record ID**.

rid	Age	Car Type	Risk
0	23	family	High
1	17	sports	High
2	43	sports	High
3	68	family	Low
4	32	truck	Low
5	20	family	High

Training Dataset

Age	Class	rid
17	High	1
20	High	5
23	High	0
32	Low	4
43	High	2
68	Low	3

Attribute List

Selecting Splitting Attribute

- For a set of records S, the **entropy**

$$E(S)=-\sum_j p_j \log p_j$$

where p_j is the relative frequency of class j in S

- The **entropy of a split** that divides S with n record into sets S_1 with n_1 records and S_2 with n_2 records is $E(S_1, S_2)= (n_1/n)\,E(S_1)+ (n_2/n)\,E(S_1)$

- Once the best split has been found,it is used to split all the attribute lists for the two child nodes.

Example of Building Phase

Attribute lists for root node

Age	Class	Tid
17	High	1
20	High	5
23	High	0
32	Low	4
43	High	2
68	Low	3

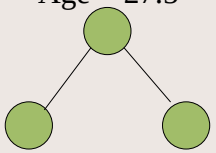
Car Type	Class	Tid
family	High	0
sports	High	1
sports	High	2
family	Low	3
truck	Low	4
family	High	5

Attribute lists for left son node

Age	Class	Tid
17	High	1
20	High	5
23	High	0

Car Type	Class	Tid
family	High	0
sports	High	1
family	High	5

Age < 27.5



Attribute lists for right son node

Age	Class	Tid
32	Low	4
43	High	2
68	Low	3

Car Type	Class	Tid
sports	High	2
family	Low	3
truck	Low	4

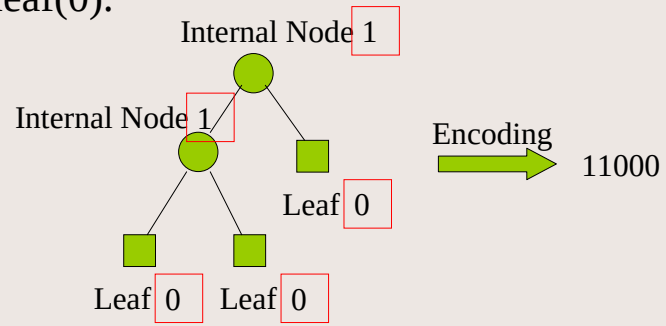
Pruning Principle

- MDL(Minimum Description Length) principle states that the "**best**" tree is the one that can be encoded using the **fewest** number of bits.

- Cost of Encoding Tree
  - -Encoding the structure of the tree
  - -Encoding each split
  - -Encoding the classes of data records in each leaf

Encoding the Structure

Using a single bit to specify whether a node of the tree is an internal node(1) or leaf(0).



Encoding the Tree(Cont')

Let a set S contain n records each belonging to one of k classes, n_i being the number of records with class i, let "α" be the number of attributes and let v be the number of distinct values for the splitting attribute in records at the node.

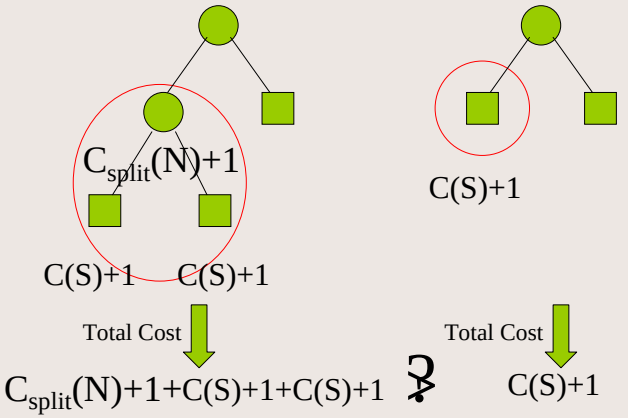
- The cost of encoding each split $C_{split}(N)$ is

  $\text{Log } \alpha + \text{Log}(v-1)$ or $\text{Log } \alpha + \text{Log}(2^v-2)$

- The cost of encoding the classes of data records

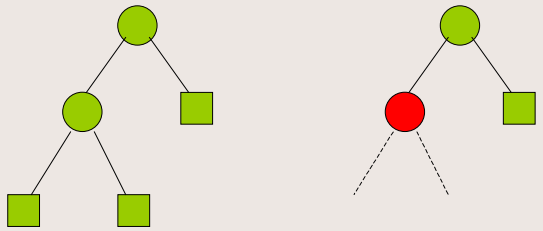
  $C(S)=\Sigma\, n_i \log(n/n_i)+(k-1)\log(n/2)+\log(pi^{k/2}/\Gamma(k/2))$

Example of Pruning Phase



PUBLIC

- PUBLIC is similar to the building procedure of SPRINT
- The difference is that periodically the **partially** built tree is pruned.



Final Tree Partial Tree

Estimating Subtree Cost

- Theorem : The cost of any subtree with s splits and rooted at Node N is at least

$$2*s+1+s*\log \alpha + \Sigma^{k}_{i=s+2} ni$$

- Using the theorem, we can compute the minimum cost for subtrees rooted at N with 0,1,2…k-1 splits and set our cost estimate to be the minimum of all these costs.

Real-life Training Set

- It comes from a big company in China and it is used to predict the activity of customers.
- Altogether the training set consists of 750 records.

Table 1 Training Dataset Used(Partial)

Type	Product	Gender	City	Age	Education	Pincome	Hincome	Married	Likability	Class
1	2	1	4	1	1	2	4	2	1	Buy
2	2	2	3	2	3	3	1	2	3	Not
1	1	1	7	1	2	1	3	1	5	Not

Experimental Results

Table 2 Execution Time (secs)

Data Set	Car	letter	satimage	shuttle	market
SPRINT	45.58	3283.2	1471	457.78	21.14
PUBLIC	33.93	2786.57	1036.34	455.15	19.03
Max Ratio	38%	18%	43%	0.60%	11%

Future Work

- How often should PUBLIC to prune the partially built tree?
- How can we estimate the subtree cost more accurately?
- Using large training sets, we may get much better performance.