A Covering-based Algorithm for Classification: PRISM

Instructor: Dr. Lisa Fan

Speaker: Xiaofei Deng

Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
E-mail: deng200x@cs.uregina.ca

CS831: Knowledge Discovery in Databases

Outline

The basic idea of ID3.

1. Greedy Algorithm.
   - Select the attribute that contributes the maximum Information Gain.
2. Inductive bias: prefers small trees over large trees.
   - A short tree but might be a wide tree.
3. Its efficiency.
   - Been proved in theory by Quinlan.
   - Works well in chess endgames.

Disadvantages of the representation of rules.

1. Difficult to manipulate for expert systems.

Extract rules about a single classification

- Need to examine the whole tree.
- Partial solution: converting Decision Trees(DT) into a set of rules.
- Problems: There're rules can't easily be represented by DT.

Example: extract rules about $C0$ from a DT

- $Rule1 : b_1 \wedge d_1 \rightarrow C0$, $Rule2 : a_3 \wedge c_1 \rightarrow C0$.
- Assume only two rules about $C0$.
- Assume no attributes common to both Rules.

Cont. (Extracting rules about *C*0)

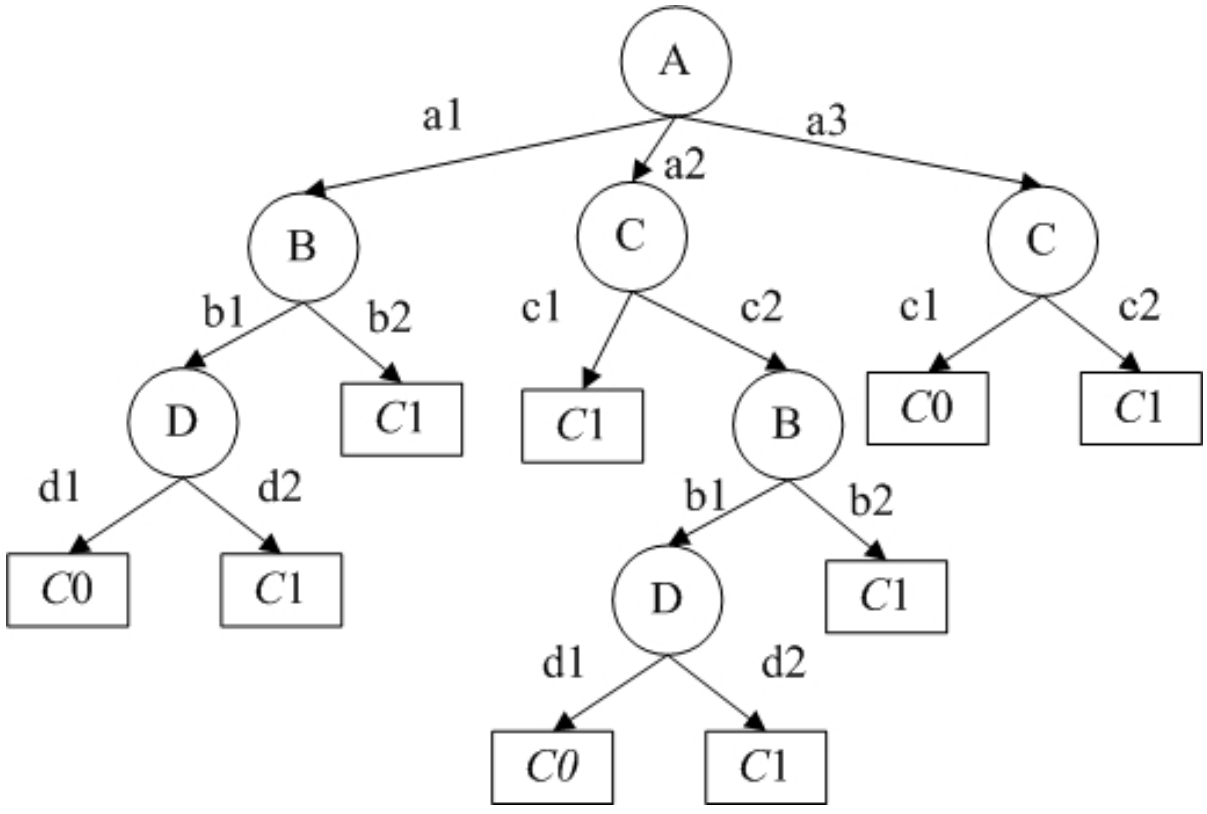


Figure: Extracting rules about *C*0 from decision tree

Cont. (Extracting rules about *C*0)

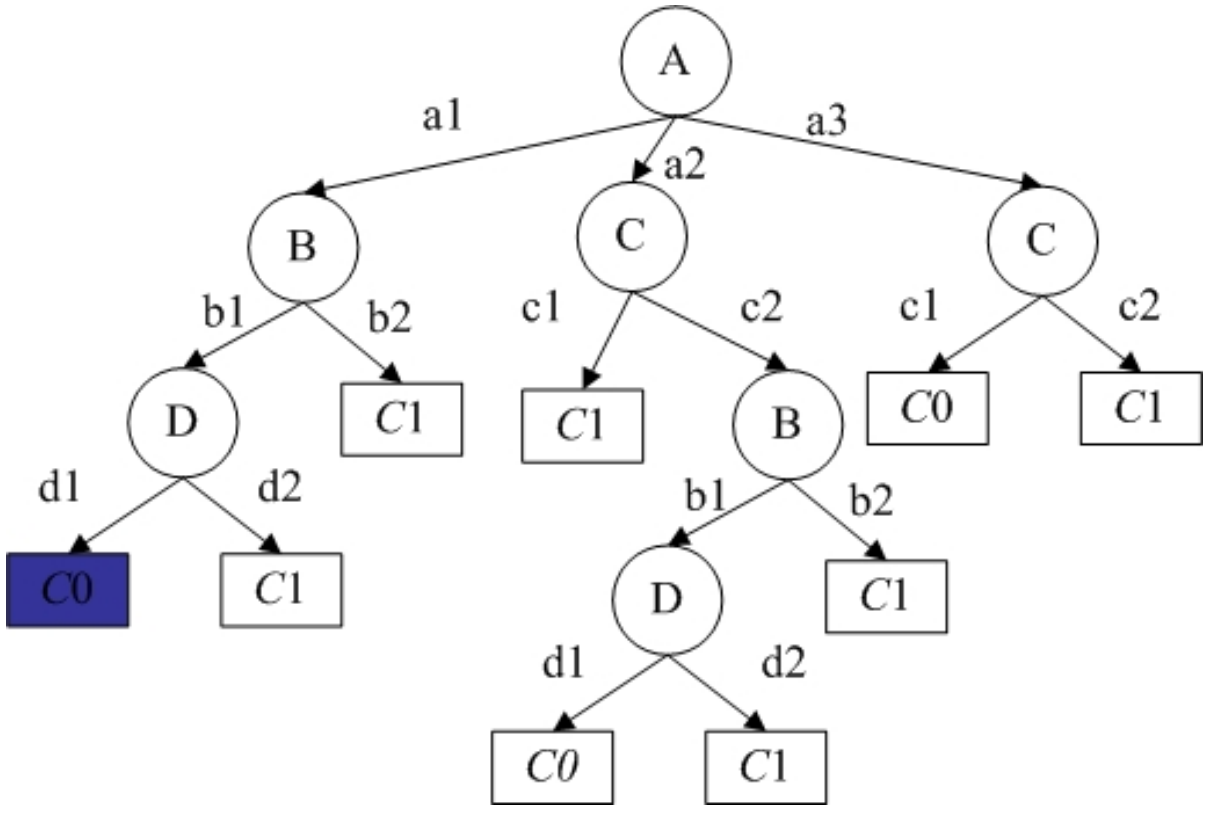


Figure: Extracting rules about *C*0 from decision tree

Cont. (Extracting rules about *C*0)

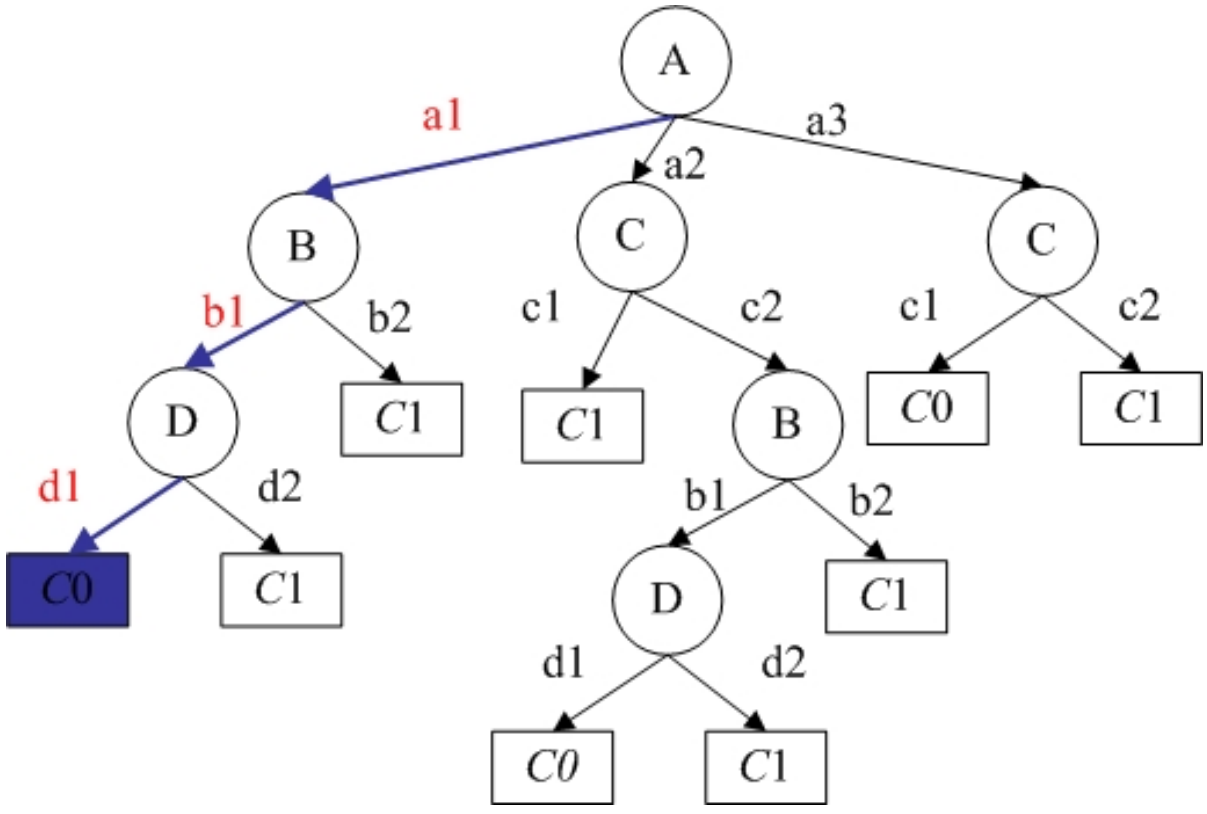


Figure: Extracting rules about *C*0 from decision tree

Cont. (Extracting rules about *C*0)

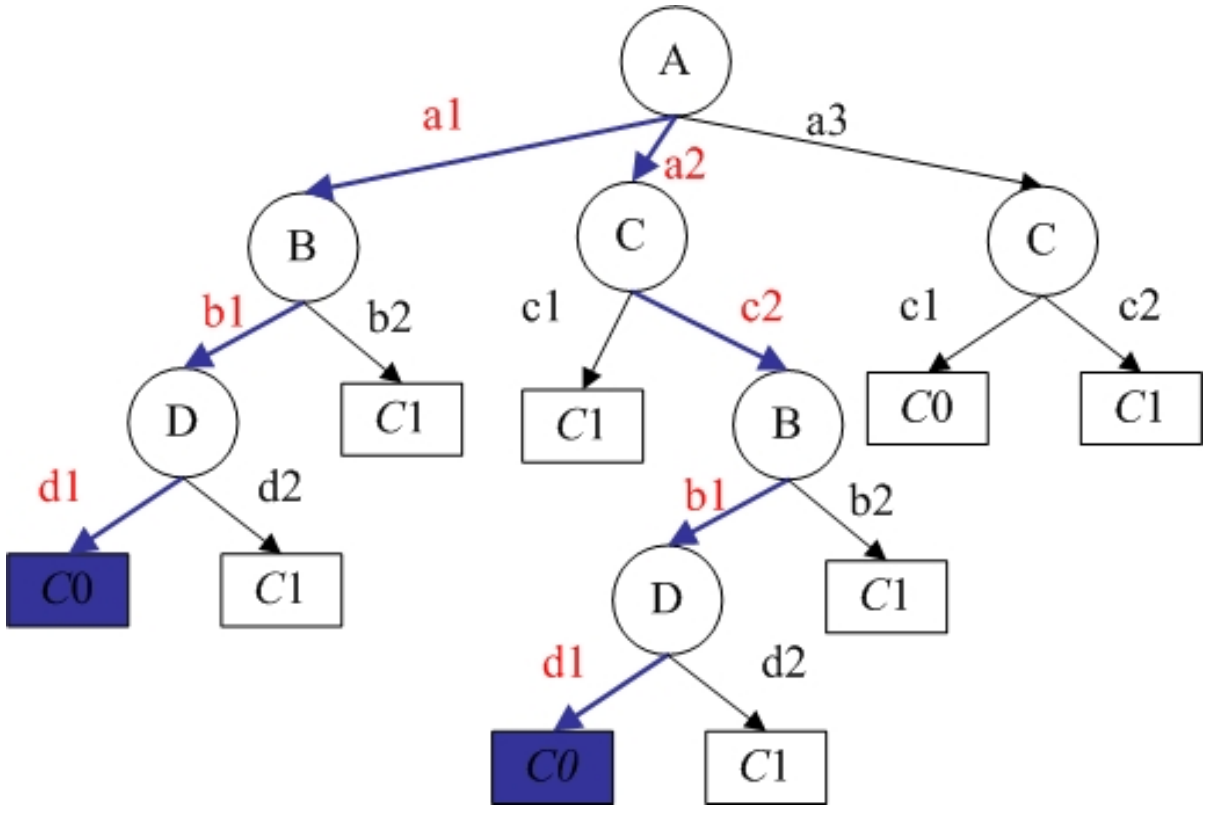


Figure: Extracting rules about *C*0 from decision tree

Cont. (Extracting rules about *C*0)

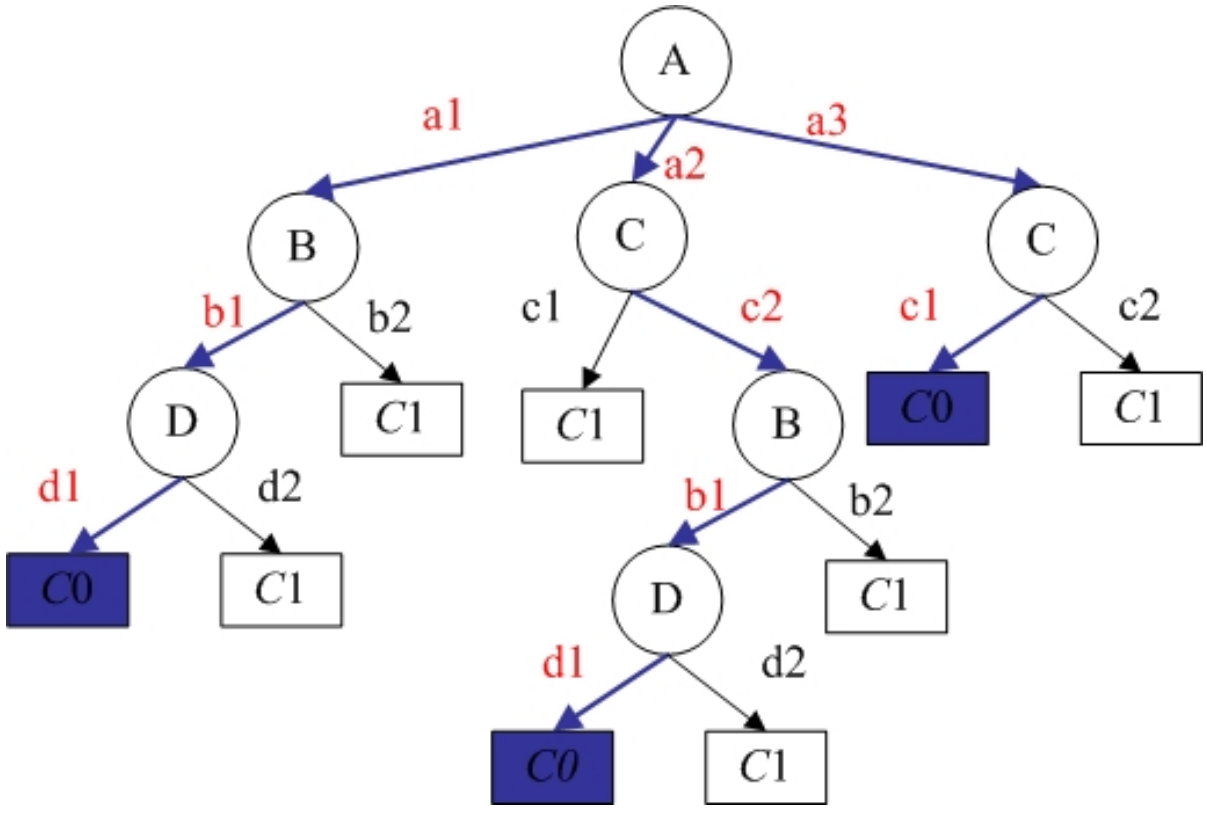


Figure: Extracting rules about *C*0 from decision tree

Cont. (Extracted rules)

Extracted Rules for Class $C0$ from DT

- $Rule1a : a_1 \wedge b_1 \wedge d_1 \rightarrow C0.$
- $Rule1b : a_2 \wedge c_2 \wedge b_1 \wedge d_1 \rightarrow C0.$
- $Rule2 : a_3 \wedge c_1 \rightarrow C0.$

Explored the whole decision tree when extracting

- Why $Rule1a, 1b$? Irrelevant attributes are added as a term to them.
- May cause serious problem, for example, a medical diagnose case which might requires an unnecessary surgery.

Information Entropy in ID3

1. The problem: ID3 Prefers an attribute which minimizes the average Entropy.

Entropy

- $$H(S) = -\sum_{i}^{n} p(C_i) log_2(c_i) bits$$
- S, n, $p(C_i)$ is the probability of occurrence of Ci.
- Entropy measures the uncertainty of current set of instances.

Why we say average Entropy?

1. Calculate the Entropy of a given set S.

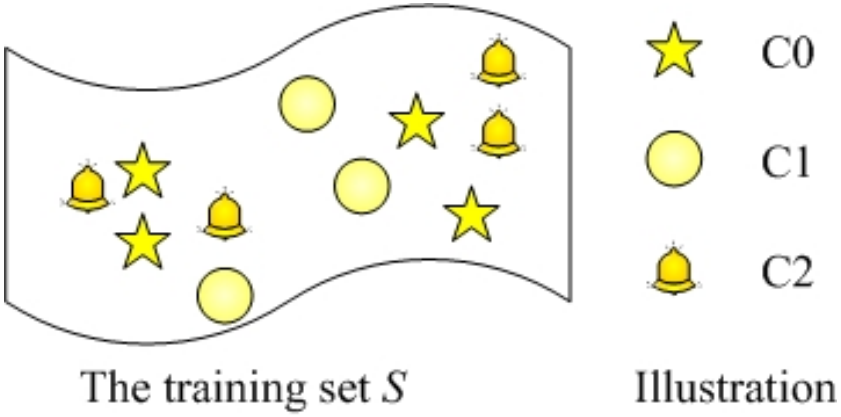


Figure: The distribution of instances of S

2. $H(S) = -p(C0)log_2p(C0) - p(C1)log_2p(C1) - p(C2)log_2p(C2)$.
3. Measures the uncertainty in Average.
   - We added them to calculate the uncertainty.
   - Using $H(S)$, means consider all three, $C0, C1, C2$.

What about the uncertainty after knowing an Attribute?

1. ID3 chooses the attribute that contributed maximum information to lower the uncertainty.
2. But, that information measures in average.

Information Gain

- $$Gain(S, A) = H(S) - \sum_i \frac{|S_{vi}|}{|S|} H(S_{vi}) bits$$
- Average entropy **Before** – **After** (knowing A).
- the second part is the info. A contributed.
- The second part measures the average information of all the branches of A.

Why the info. contributed by an attribute measures in average?

1. When choose attribute A ($Gain(S, A)$ has max. value).
2. A partitions S into three branches,S_{v1}, S_{v2}, S_{v3}.

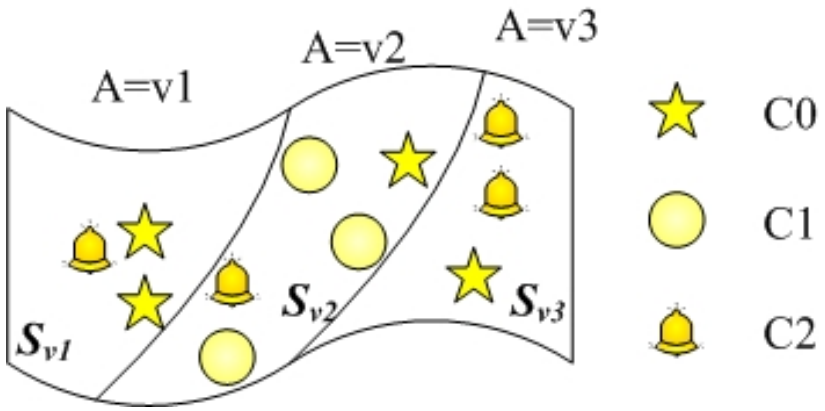
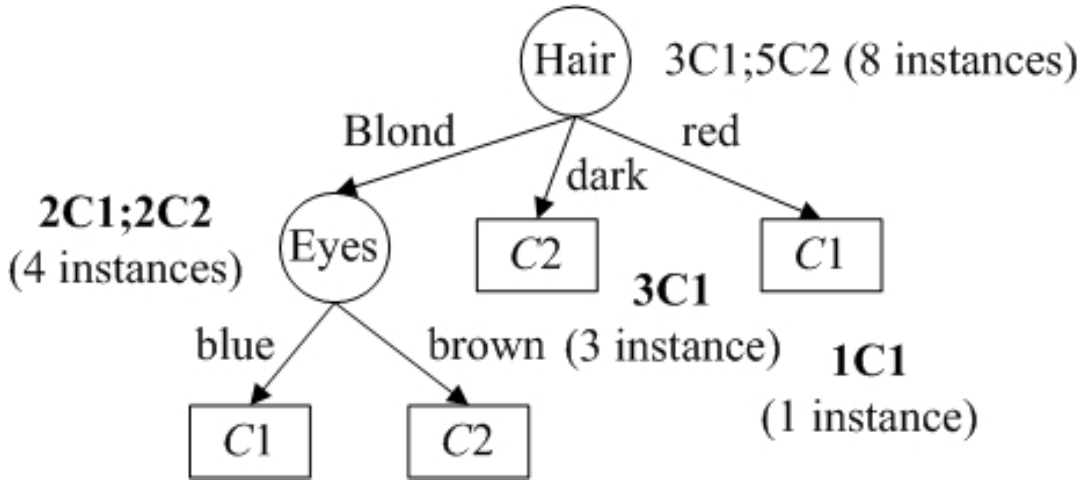


Figure: The training set S is partitioned by A

3. $$\sum_i \frac{|S_{vi}|}{|S|} H(S_{vi}) bits = \frac{|S_{v1}|}{|S|} Entropy(Branch\, v1)$$
$$+ \frac{|S_{v2}|}{|S|} Entropy(Branch\, v2) + \frac{|S_{v3}|}{|S|} Entropy(Branch\, v3)$$

Average dose not mean Good

1. An example: sometimes it would be worse for a branch



2. The average uncertainty of A is low.
   - $$\sum_{1}^{3} \frac{|S_{vi}|}{|S|} H(S_{vi}) = 0.25 bits$$

3. Uncertainty some branches of A is low, some rather high
   - Branch $Hair = Blond$ is 0.5. high
   - Branch $Hair = dark$, $Hair = red$ is 0.(low)

A short summary of the inner weakness of ID3

ID3

- ID3 is attribute oriented.
- Selecting an attribute, then all the sub-branches are consider in average.
- ID3 measures the average information entropy.
- Average doesn't mean good to each rule.

ID3 doesn't consider following cases

- An attribute might be highly **relevant** to only one classification and **irrelevant** to the others.
- Sometimes only one value of the attribute is **relevant**.

How does PRISM fix this problem?

The strategy of PRISM

- A branch could be considered as an attribute-value pair.
- Consider the relevance between an attribute-value pair and the specific classification.
- Choose the attribute-value pair that contributes maximum information as the term of a rule for one specific classification.

An Information theoretic approach: PRISM

1. The task of PRISM.

Find the α_x that contributes maximum Information about *Ci*.

- An attribute-value pair, α_x.
- A specific classification, *Ci*.

2. The amount of Information about occurrence of *Ci* given α_x is told:
$I(Ci, \alpha_x)$

$$= log_2\left(\frac{Probability\ of\ occurence\ of\ Ci\ after\ knowing\ \alpha_x}{Probability\ of\ occurence\ of\ Ci\ before\ knowing\ \alpha_x}\right) bits$$

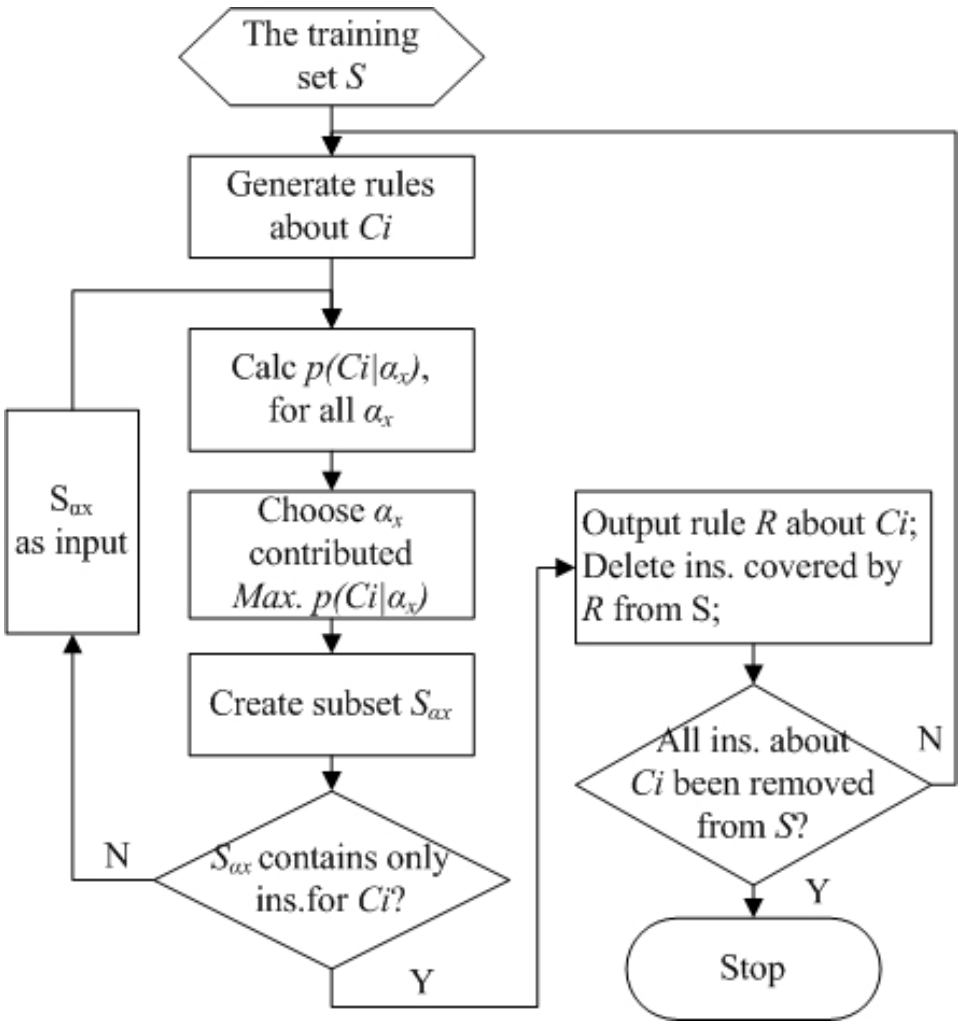
$$= log_2\left(\frac{p(Ci|\alpha_x)}{p(Ci)}\right) bits$$

Cont.

1. $I(Ci, \alpha_x) = log_2(\frac{p(Ci|\alpha_x)}{p(Ci)})bits$
2. $p(Ci|\alpha_x) = \frac{Number\ of\ instances\ labeled\ Ci}{|S_{\alpha_x}|}$
   - The **After**.
   - The probability of occurrence of Ci in S_{α_x}.
   - S_{α_x} is the subset of instances contain α_x.
3. $p(Ci) = \frac{Number\ of\ instances\ labeled\ Ci}{|S|}$
   - The **Before**.
   - The probability of occurrence of Ci in S.
   - For all the α_x, it's the same.
   - Thus, we only calculate the $p(Ci|\alpha_x)$.

PRISM algorithm: the basic steps

1. Steps for generating rules about *Ci*, like *C*1.



Cont.(steps in detail)

1. Calculate the probability of occurrence, $p(Ci|\alpha_x)$, of the classification *Ci* for each attribute-value pair.
2. Select the attribute-value pair α_x for which $p(Ci|\alpha_x)$ is maximum, and create a subset, S_{α_x}, that contains instances with α_x.
3. Repeat step 1 and 2 for the subset, until it contains only instances for classification *Ci*. The induced rule is a conjunction of all the attribute-value pairs used in creating the subset.
4. remove all instances covered by this rule from the training set *S*.
5. Repeat Steps 1-4 until all instances of class *Ci* have been removed.

Note. (For those steps)

1. $p(Ci|\alpha_x)$ measures the contribution of α_x.
2. Trying to find all rules about one specific classification Ci.

Rules about Class $C1$

- $Rule1 : b_1 \wedge d_1 \rightarrow C1$.
- $Rule2 : a_3 \wedge c_1 \rightarrow C1$.

3. A rule is the conjunction of attribute-value pairs.

Generating a rule about Class $C1$

- $\alpha_1 : Hair = Blond$. (1st attribute-value pair, term)
- $\alpha_2 : Eyes = Blue$. (2nd pair, term)
- $Rule1 : (Hair = Blond \wedge Eyes = Blue) \rightarrow C1$

Note. (For those steps)

1. $p(Ci|\alpha_x)$ measures the contribution of α_x.
2. Trying to find all rules about one specific classification Ci.

Rules about Class $C1$

- $Rule1 : b_1 \wedge d_1 \rightarrow C1.$
- $Rule2 : a_3 \wedge c_1 \rightarrow C1.$

Then $C2, \ldots$

- $Rule3 : p_3 \wedge q_7 \rightarrow C2.$
- $Rule4 : k_2 \wedge t_5 \rightarrow C2.$

3. A rule is the conjunction of attribute-value pairs.

Generating a rule about Class $C1$

- $\alpha_1 : Hair = Blond$. (1st attribute-value pair, term)
- $\alpha_2 : Eyes = Blue$. (2nd pair, term)
- $Rule1 : (Hair = Blond \wedge Eyes = Blue) \rightarrow C1$

An example for calculation

1. Current training set $S = \{1, 2, 3, 4, 5, 6, 7, 8\}$.

Object	Height	Hair	Eyes	Class
O1	short	blond	blue	C1
O2	short	blond	brown	C2
O3	tall	red	blue	C1
O4	tall	dark	blue	C2
O5	tall	dark	blue	C2
O6	tall	blond	blue	C1
O7	tall	dark	brown	C2
O8	short	blond	brown	C2

Generate rules for $C1$

1. Find 1st rule about $C1$ ($\rightarrow C1$)

Generate rules for $C1$

1. Find 1st rule about $C1$ ($\rightarrow C1$)
2. Calculate all the $p(C1|\alpha_x)$ for all α_x

α_x	$C1$ (instances)	S_{α_x}	$p(C1\|\alpha_x)$
Height=short	{1}	{1,2,8}	1/3=0.333
Height=tall	{3,6}	{3,4,5,6,7}	2/5=0.4
Hair=blond	{1,6}	{1,2,6,8}	2/4=0.5
Hair=red	{3}	{3}	1/1=1
Hair=dark	{}	{4,5,7}	0
Eyes=blue	{1,3,6}	{1,3,4,5,6}	3/5=0.6
Eyes=brown	{}	{2,7,8}	0

Figure: Probability of occurrence of $C1$ with each pair

Calculate $p(C1|Hair = blond)$

1. Probability of occurrence of $C1$ with α_x : $Hair = blond$.

Calculate $p(C1|Hair = blond)$

1. Probability of occurrence of $C1$ with α_x : $Hair = blond$.

Object	Height	Hair	Eyes	Class
O1	short	blond	blue	C1
O2	short	blond	brown	C2
O6	tall	blond	blue	C1
O8	short	blond	brown	C2

2. $p(C1|\alpha_x) = p(C1|Hair = blond) = \frac{|\{1,6\}|}{|\{1,2,6,8\}|} = \frac{2}{4} = 0.5$.

Output the *Rule*1

1. Choose α_x : *Hair* $=$ *red* as the first term for *Rule*1 : (*Hair* $=$ *red*) $\wedge$ (. . .) $\rightarrow$ *C*1.

Output the *Rule*1

1. Choose α_x : $Hair = red$ as the first term for $Rule1 : (Hair = red) \wedge (\ldots) \rightarrow C1$.
2. Create subset $S_{\alpha_x} = S_{Hair=red} = \{3\}$

Output the *Rule*1

1. Choose α_x : *Hair* $=$ *red* as the first term for *Rule*1 : (*Hair* $=$ *red*) $\wedge$ (...) $\rightarrow$ *C*1.
2. Create subset $S_{\alpha_x} = S_{Hair=red} = \{3\}$
3. $S_{Hair=red} = \{3\}$ contains only instance *Object*3 labeled by *C*1.

Output the *Rule*1

1. Choose α_x : $Hair = red$ as the first term for $Rule1 : (Hair = red) \wedge (\ldots) \rightarrow C1$.
2. Create subset $S_{\alpha_x} = S_{Hair=red} = \{3\}$
3. $S_{Hair=red} = \{3\}$ contains only instance *Object*3 labeled by *C*1.
4. Output the $Rule1 : (Hair = red) \rightarrow C1$.

Delete *Object*3 from the training set

1. Delete *Object*3 from S, thus $S = \{1, 2, 4, 5, 6, 7, 8\}$.

Delete *Object*3 from the training set

1. Delete *Object*3 from S, thus $S = \{1, 2, 4, 5, 6, 7, 8\}$.
2. Current training set $S = \{1, 2, 4, 5, 6, 7, 8\}$.

Object	Height	Hair	Eyes	Class
O1	short	blond	blue	C1
O2	short	blond	brown	C2
~~O3~~	~~tall~~	~~red~~	~~blue~~	~~C1~~
O4	tall	dark	blue	C2
O5	tall	dark	blue	C2
O6	tall	blond	blue	C1
O7	tall	dark	brown	C2
O8	short	blond	brown	C2

Repeat to find the *Rule2* about *C*1

1. Recalculate the $p(C1|\alpha_x)$ for all α_x.

α_x	$C1$ (instances)	S_{α_x}	$p(C1\,|\,\alpha_x)$
Height=short	{1}	{1,2,8}	1/3=0.333
Height=tall	{6}	{4,5,6,7}	1/4=0.25
Hair=blond	{1,6}	{1,2,6,8}	2/4=0.5
Hair=dark	{}	{4,5,7}	0
Eyes=blue	{1,6}	{1,4,5,6}	2/4=0.5
Eyes=brown	{}	{2,7,8}	0

Figure: Selecting the first term of *Rule2* about *C*1

2. *Hair* = *blond*, *Eyes* = *blue* have the equal value.
3. Choose *Hair* = *blond* as 1st term for *Rule2*.

The second term of *Rule2* about *C1*

1. Create the subset $S_{\alpha_x} = S_{Hair=blond} = \{1, 2, 6, 8\}$
2. *Object2* and *Object8* are labeled with *C2*.
3. Take $S_{\alpha_x} = S_{Hair=blond} = \{1, 2, 6, 8\}$ as the current set. Trying to find second term.

Table The subset $S_{\alpha_x} = S_{Hair=blond}$

Object	Height	Hair	Eyes	Class
O1	short	blond	blue	C1
O2	short	blond	brown	C2
O6	tall	blond	blue	C1
O8	short	blond	brown	C2

The second term of *Rule2* about *C*1

1. Create the subset $S_{\alpha_x} = S_{Hair=blond} = \{1, 2, 6, 8\}$
2. *Object*2 and *Object*8 are labeled with *C*2.
3. Take $S_{\alpha_x} = S_{Hair=blond} = \{1, 2, 6, 8\}$ as the current set. Trying to find second term.

α_x	$C1$ (instances)	S_{α_x}	$p(C1 \mid \alpha_x)$
Height=short	{1}	{1,2,8}	1/3=0.333
Height=tall	{1}	{1}	1/1=1
Eyes=blue	{1,6}	{1, 6}	2/2=1
Eyes=brown	{}	{2,8}	0

Cont.

1. Choose the *Eyes* $=$ *blue* as the second term (consistent).

Cont.

1. Choose the $Eyes = blue$ as the second term (consistent).
2. Create subset $S_{\alpha_x'} = S_{Hair=blond \wedge Eyes=blue} = \{1, 6\}$.

Cont.

1. Choose the $Eyes = blue$ as the second term (consistent).
2. Create subset $S_{\alpha_x'} = S_{Hair=blond \wedge Eyes=blue} = \{1, 6\}$.
3. $\{1, 6\}$ are all labeled with $C1$, output $Rule2$.

Cont.

1. Choose the $Eyes = blue$ as the second term (consistent).
2. Create subset $S_{\alpha_x'} = S_{Hair=blond \wedge Eyes=blue} = \{1, 6\}$.
3. $\{1, 6\}$ are all labeled with $C1$, output $Rule2$.
4. $Rule2 : (Hair = blond \wedge Eyes = blue) \rightarrow C1$.

Cont.

1. Choose the $Eyes = blue$ as the second term (consistent).
2. Create subset $S_{\alpha_x'} = S_{Hair=blond \wedge Eyes=blue} = \{1, 6\}$.
3. $\{1, 6\}$ are all labeled with $C1$, output $Rule2$.
4. $Rule2 : (Hair = blond \wedge Eyes = blue) \rightarrow C1$.
5. Delete Object $1, 6$ from current training set.

Cont.

1. Choose the $Eyes = blue$ as the second term (consistent).
2. Create subset $S_{\alpha_x'} = S_{Hair=blond \wedge Eyes=blue} = \{1, 6\}$.
3. $\{1, 6\}$ are all labeled with $C1$, output $Rule2$.
4. $Rule2 : (Hair = blond \wedge Eyes = blue) \rightarrow C1$.
5. Delete Object $1, 6$ from current training set.
6. No others instances labeled with $C1$, stop.

Cont.

1. Choose the $Eyes = blue$ as the second term (consistent).
2. Create subset $S_{\alpha_x'} = S_{Hair=blond \wedge Eyes=blue} = \{1, 6\}$.
3. $\{1, 6\}$ are all labeled with $C1$, output $Rule2$.
4. $Rule2 : (Hair = blond \wedge Eyes = blue) \rightarrow C1$.
5. Delete Object $1, 6$ from current training set.
6. No others instances labeled with $C1$, stop.
7. Repeat above steps for $C2$.

The results by PRISM and ID3

Results by PRISM

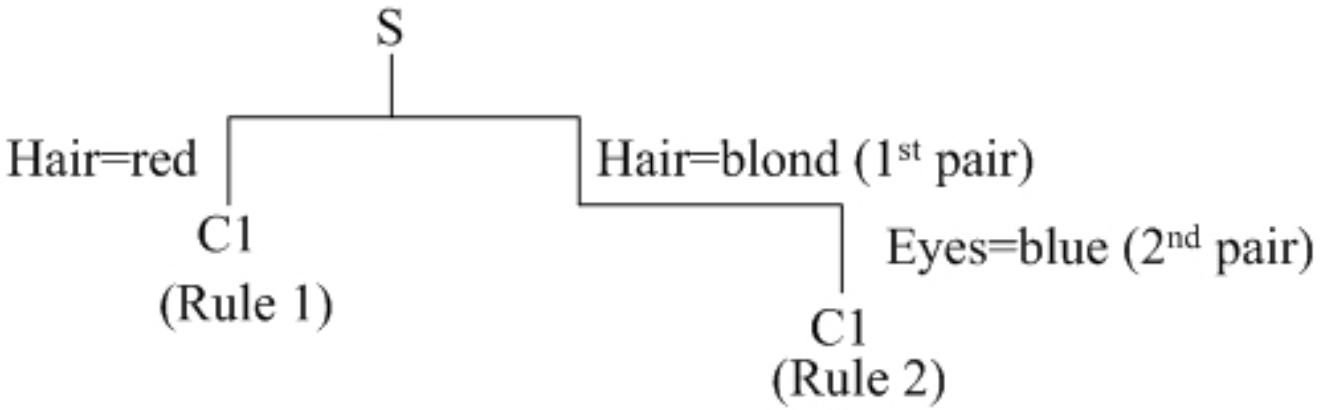
- $(Hair = red) \rightarrow C1$.
- $(Hair = blond \wedge Eyes = blue) \rightarrow C1)$.
- $(Eyes = brown) \rightarrow C2$.
- $(Hair = dark) \rightarrow C2$.

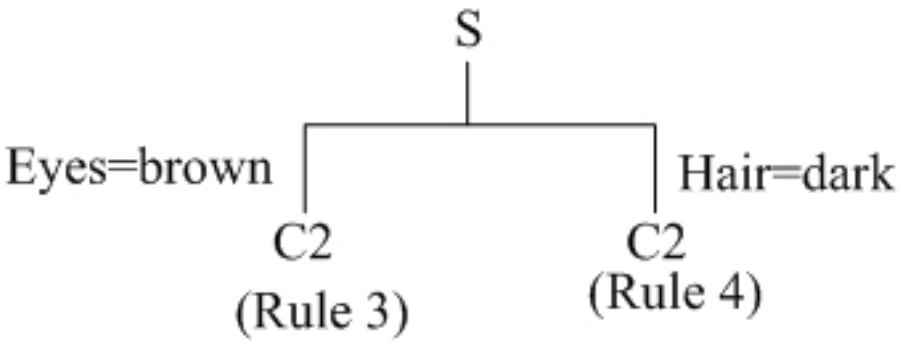
Results by ID3

- $(Hair = red) \rightarrow C1$.
- $(Hair = blond \wedge Eyes = blue) \rightarrow C1)$.
- $(Hair = blond \wedge Eyes = brown) \rightarrow C2$.
- $(Hair = dark) \rightarrow C2$.

Cont.

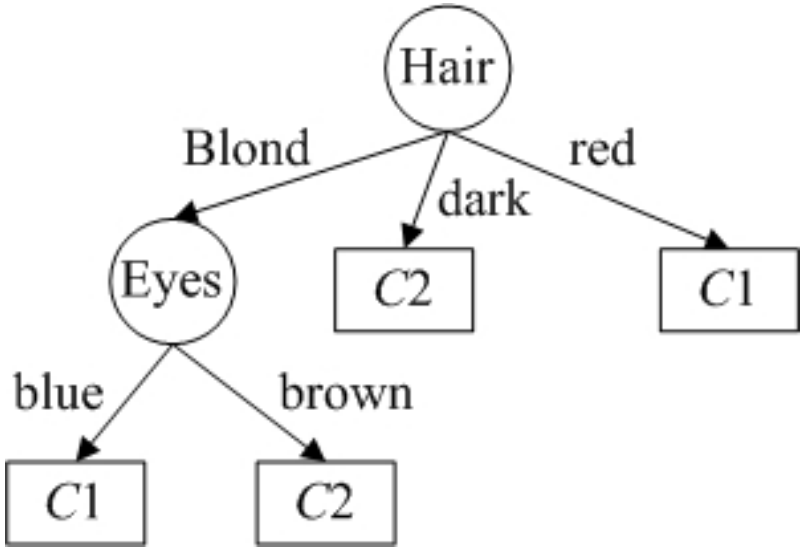
1. 'Decision Tree' by PRISM





Cont.

1. Decision Tree by ID3



Summary

ID3

- Greedy algorithm.
- Measures average information an attribute contributed.
- Attribute-oriented.
- Rules might contain irrelevant attributes.

PRISM

- Greedy algorithm.
- Measures the attribute-value pair in determination of the classification.
- Attribute-value-oriented.
- More general and less rules.

Q.&A.

Any questions?