SLIQ Algorithm
for disk resident data

SLIQ

- SLIQ is a decision tree classifier that can handle both **numerical** and **categorical** attributes

- Uses a pre-sorting technique in the tree growing phase

- Suitable for classification of **large disk-resident** datasets

Issues

- There are two major, critical performance, issues in the tree-growth phase:
  - How to find split points
  - How to partition the data

- The well-known decision tree classifiers:
  - Grow trees depth-first
  - Repeatedly sort the data at every node

- SLIQ:
  - Replace this repeated sorting with one-time sort
  - Use new a data structure called class-list
  - Class-list must remain memory resident at all times

Some Data

rid	age	salary	marital	car
1	30	60	single	sports
2	25	20	single	mini
3	40	80	married	van
4	45	100	single	luxury
5	60	150	married	luxury
6	35	120	single	sports
7	50	70	married	van
8	55	90	single	sports
9	65	30	married	mini
10	70	200	single	luxury

SLIQ - Attribute Lists

rid	age
1	30
2	25
3	40
4	45
5	60
6	35
7	50
8	55
9	65
10	70

rid	salary
1	60
2	20
3	80
4	100
5	150
6	120
7	70
8	90
9	30
10	200

rid	marital
1	single
2	single
3	married
4	single
5	married
6	single
7	married
8	single
9	married
10	single

These are projections on (rid, attribute).

SLIQ - Sort Numeric, Group Categorical

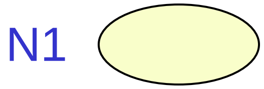
rid	age
2	25
1	30
6	35
3	40
4	45
7	50
8	55
5	60
9	65
10	70

rid	salary
2	20
9	30
1	60
7	70
3	80
8	90
4	100
6	120
5	150
10	200

rid	marital
3	married
5	married
7	married
9	married
1	single
2	single
4	single
6	single
8	single
10	single

SLIQ - Class List

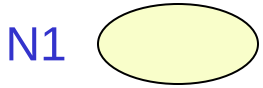
rid	car	LEAF
1	sports	N1
2	mini	N1
3	van	N1
4	luxury	N1
5	luxury	N1
6	sports	N1
7	van	N1
8	sports	N1
9	mini	N1
10	luxury	N1



SLIQ - Histograms

rid	age
2	25
1	30
6	35
3	40
4	45
7	50
8	55
5	60
9	65
10	70

rid	car	LEAF
1	sports	N1
2	mini	N1
3	van	N1
4	luxury	N1
5	luxury	N1
6	sports	N1
7	van	N1
8	sports	N1
9	mini	N1
10	luxury	N1



	sports	mini	van	luxury
L	0	0	0	0
R	3	2	2	3

age≤25 ?

	sports	mini	van	luxury
L				
R				

age≤30 ?

	sports	mini	van	luxury
L				
R				

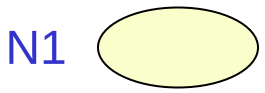
...

Evaluate each split,
using Entropy or GINI.

SLIQ - Histograms

rid	age
2	25
1	30
6	35
3	40
4	45
7	50
8	55
5	60
9	65
10	70

rid	car	LEAF
1	sports	N1
2	mini	N1
3	van	N1
4	luxury	N1
5	luxury	N1
6	sports	N1
7	van	N1
8	sports	N1
9	mini	N1
10	luxury	N1



	sports	mini	van	luxury
L	0	0	0	0
R	3	2	2	3

age≤25

	sports	mini	van	luxury
L	0	1	0	0
R	3	1	2	3

age≤30

	sports	mini	van	luxury
L	1	1	0	0
R	2	1	2	3

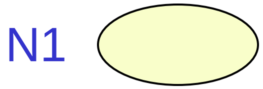
...

Evaluate each split,
using Entropy or GINI.

SLIQ - Histograms

rid	salary
2	20
9	30
1	60
7	70
3	80
8	90
4	100
6	120
5	150
10	200

rid	car	LEAF
1	sports	N1
2	mini	N1
3	van	N1
4	luxury	N1
5	luxury	N1
6	sports	N1
7	van	N1
8	sports	N1
9	mini	N1
10	luxury	N1



	sports	mini	van	luxury
L	0	0	0	0
R	3	2	2	3

salary≤20

	sports	mini	van	luxury
L	0	1	0	0
R	3	1	2	3

salary≤30

	sports	mini	van	luxury
L	0	2	0	0
R	3	0	2	3

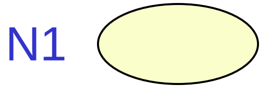
...

Evaluate each split,
using Entropy or GINI.

SLIQ - Histograms

rid	marital
3	married
5	married
7	married
9	married
1	single
2	single
4	single
6	single
8	single
10	single

rid	car	LEAF
1	sports	N1
2	mini	N1
3	van	N1
4	luxury	N1
5	luxury	N1
6	sports	N1
7	van	N1
8	sports	N1
9	mini	N1
10	luxury	N1



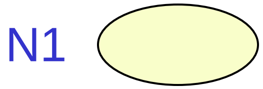
	sports	mini	van	luxury
Married				
Single				

Evaluate each split,
using Entropy or GINI.

SLIQ - Histograms

rid	marital
3	married
5	married
7	married
9	married
1	single
2	single
4	single
6	single
8	single
10	single

rid	car	LEAF
1	sports	N1
2	mini	N1
3	van	N1
4	luxury	N1
5	luxury	N1
6	sports	N1
7	van	N1
8	sports	N1
9	mini	N1
10	luxury	N1

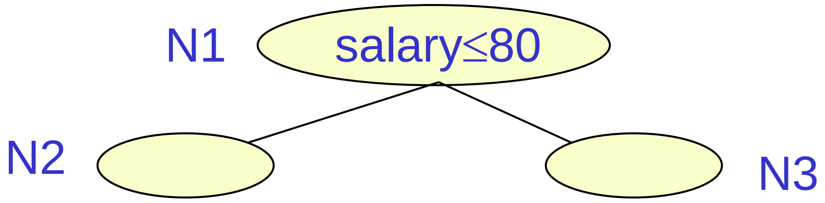


	sports	mini	van	luxury
Married	0	1	2	1
Single	3	1	0	2

Evaluate each split,
using Entropy or GINI.

SLIQ - Perform split(s)

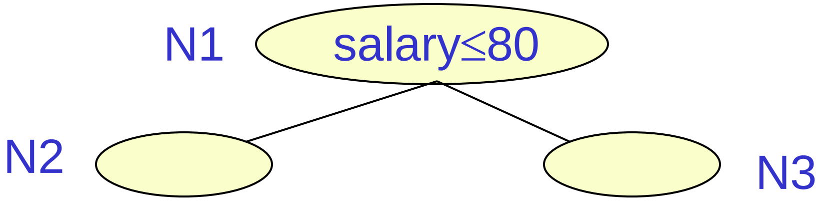
rid	car	LEAF
1	sports	N1
2	mini	N1
3	van	N1
4	luxury	N1
5	luxury	N1
6	sports	N1
7	van	N1
8	sports	N1
9	mini	N1
10	luxury	N1



SLIQ - Update Class List

rid	salary
2	20
9	30
1	60
7	70
3	80
8	90
4	100
6	120
5	150
10	200

rid	car	LEAF
1	sports	N1
2	mini	N1
3	van	N1
4	luxury	N1
5	luxury	N1
6	sports	N1
7	van	N1
8	sports	N1
9	mini	N1
10	luxury	N1

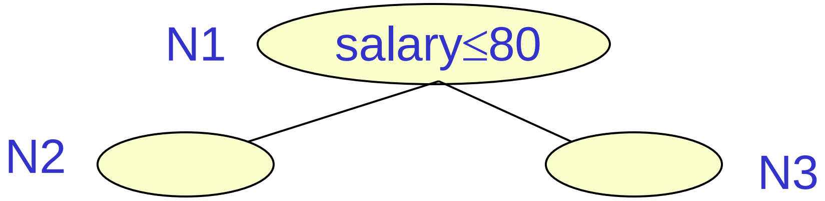


Read salary list again.

SLIQ - Update Class List

rid	salary
2	20
9	30
1	60
7	70
3	80
8	90
4	100
6	120
5	150
10	200

rid	car	LEAF
1	sports	N2
2	mini	N2
3	van	N2
4	luxury	N3
5	luxury	N3
6	sports	N3
7	van	N2
8	sports	N3
9	mini	N2
10	luxury	N3



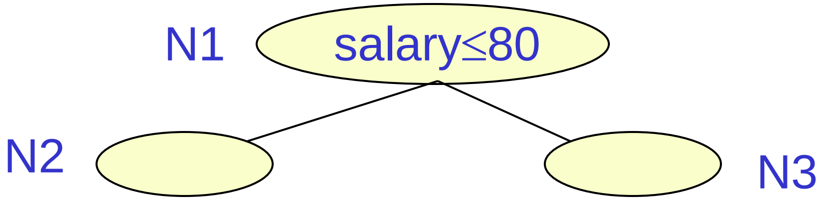
Read salary list again.

SLIQ - Histograms

rid	age
2	25
1	30
6	35
3	40
4	45
7	50
8	55
5	60
9	65
10	70

rid	car	LEAF
1	sports	N2
2	mini	N2
3	van	N2
4	luxury	N3
5	luxury	N3
6	sports	N3
7	van	N2
8	sports	N3
9	mini	N2
10	luxury	N3

Evaluate each split,
using GINI or Entropy.



N2

	sports	mini	van	luxury
L	0	0	0	0
R	1	2	2	0

N3

	sports	mini	van	luxury
L	0	0	0	0
R	2	0	0	3

age≤25 ?

N2

	sports	mini	van	luxury
L				
R				

N3

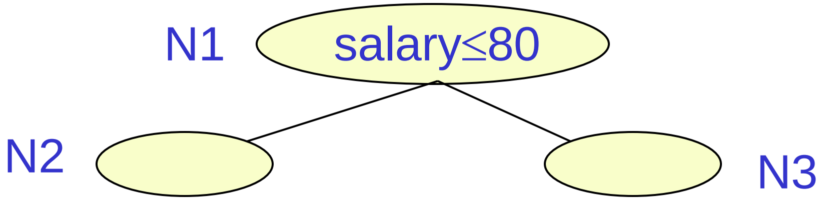
	sports	mini	van	luxury
L				
R				

...

SLIQ - Histograms

rid	age
2	25
1	30
6	35
3	40
4	45
7	50
8	55
5	60
9	65
10	70

rid	car	LEAF
1	sports	N2
2	mini	N2
3	van	N3
4	luxury	N3
5	luxury	N3
6	sports	N3
7	van	N3
8	sports	N3
9	mini	N2
10	luxury	N3



N2

	sports	mini	van	luxury
L	0	0	0	0
R	1	1	3	0

N3

	sports	mini	van	luxury
L	0	0	0	0
R	2	0	0	3

age≤25

N2

	sports	mini	van	luxury
L	0	1	0	0
R	1	0	3	0

N3

	sports	mini	van	luxury
L	0	0	0	0
R	2	0	0	3

...

Evaluate each split,
using GINI or Entropy.

SLIQ - Pseudocode

- Split evaluation:

```
EvaluateSplits()
    for each numeric attribute $A$ do
       for each value $v$ in the attribute list do
            find the corresponding entry in the class list, and
               hence the corresponding class and the leaf node $N_i$
            update the class histograms for leaf $N_i$
            compute splitting score for test ($A \le v$) for $N_i$

    for each categorical attribute $A$ do
       for each leaf of the tree do
            find subset of $A$ with best split

    return set nodes_to_split
```

SLIQ - Pseudocode

- Update class list

```
UpdateLabels()
  for each attribute $A$ used in a split do
        traverse attribute list of $A$
        for each value $v$ in the attribute list do
              find the corresponding entry in the class list (say $e$)
              find the new node $n$ to which $e$ belongs
                    by applying the splitting test $A \leq v$ at the node
                          referenced from e
              update the reference in $e$ to the child
                    corresponding to node $n$
```

SLIQ - requirement

- Class-list must remain memory resident at all time!
  - Although not a big problem with today's memories, still there might be cases where this is a bottleneck.