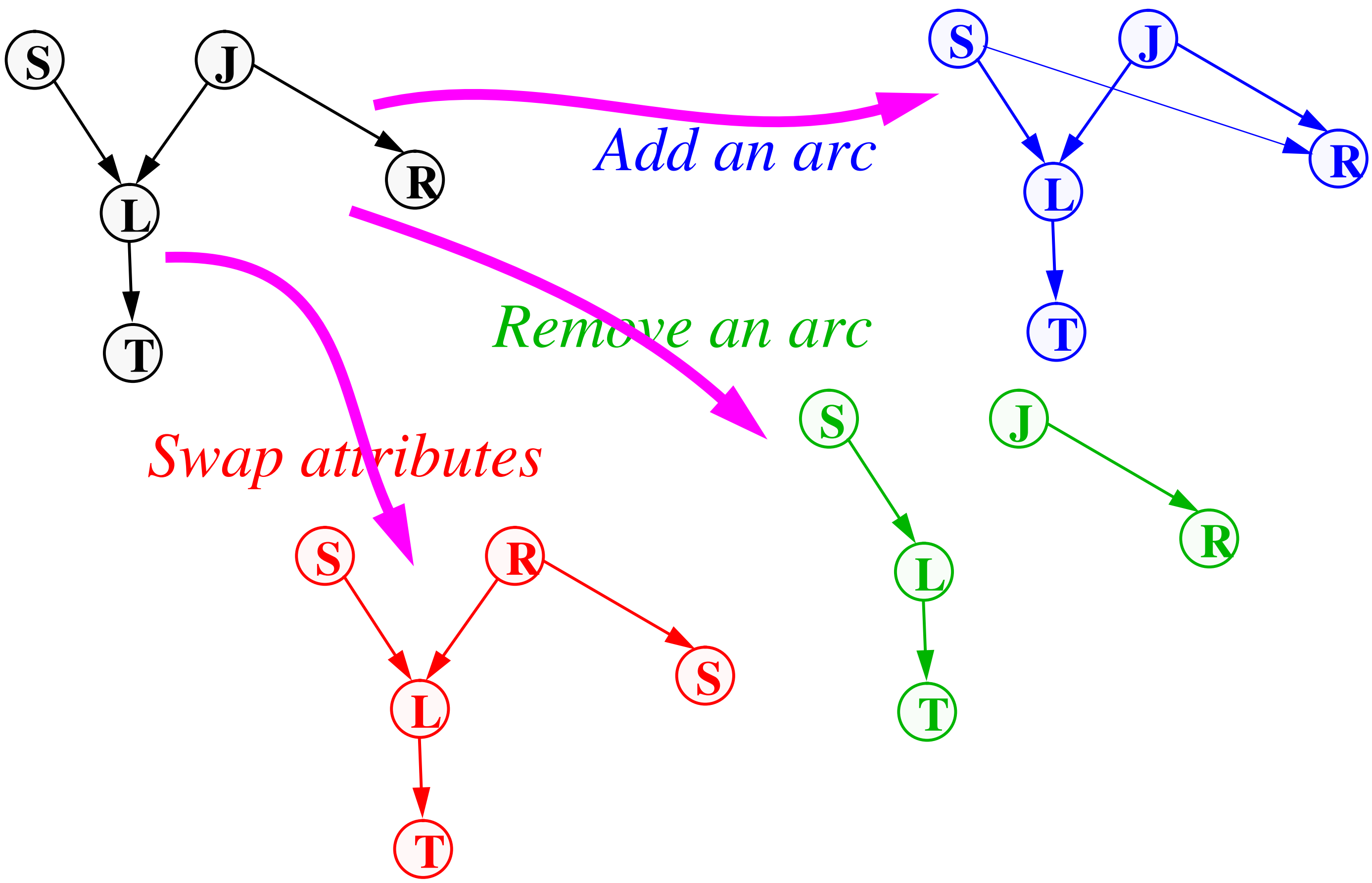
Searching for a Bayes Net structure



Simulated annealing with random restarts .

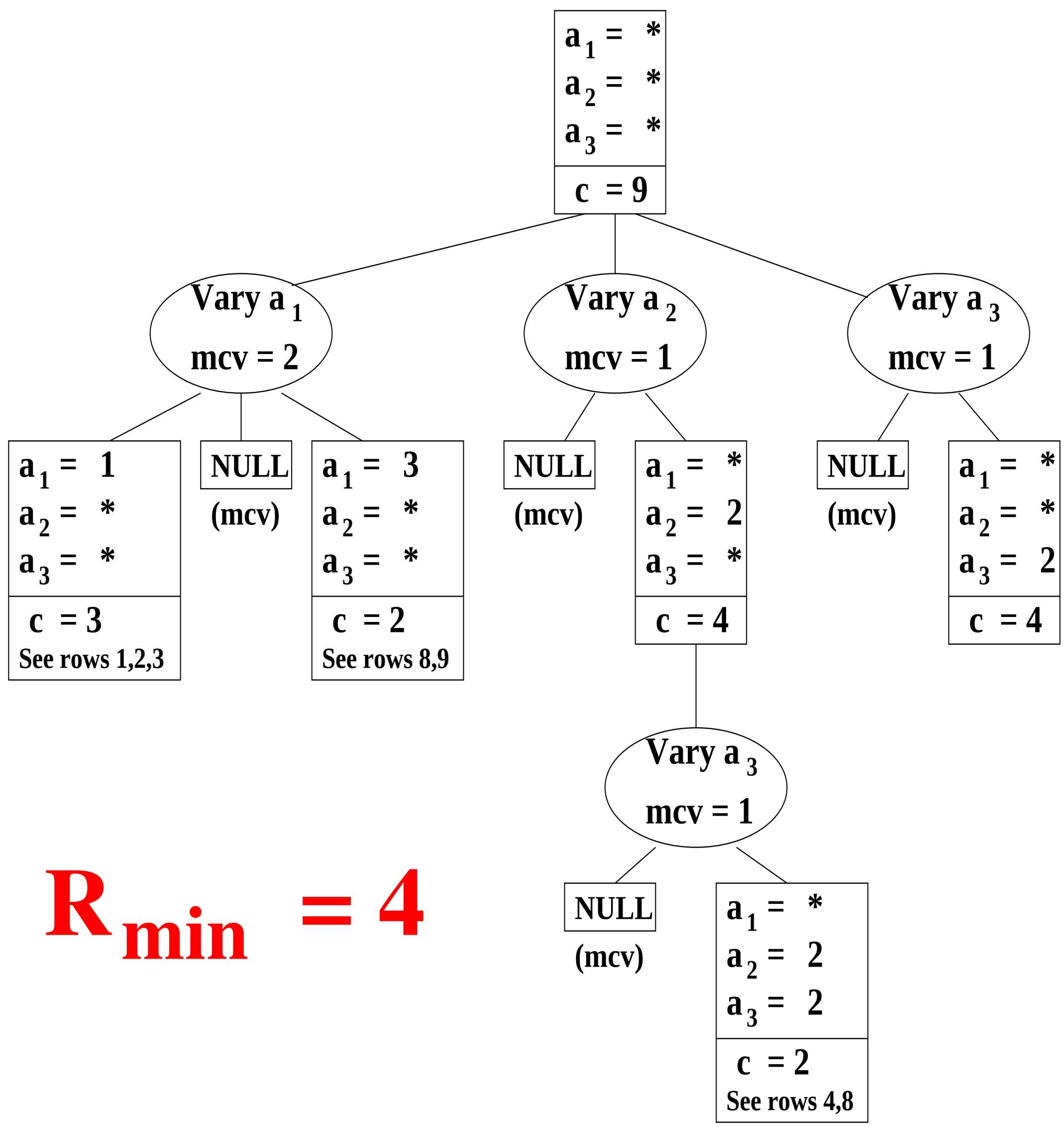
Eac h c hange requires re-evaluation of one or more contingency tables .

Datasets Used

Name	R = Num Records	N = Num Atts	Explain
ADUL T1	15060	15	The small ``Adult Income'' dataset placed in the UCI repository by Ron Kohavi. Contains census data related to job;wealth;and nationality . Attribute arities range from 2 to 41.
ADUL T2	30162	15	The same kinds of records as above but with different data.
ADUL T3	45222	15	ADUL T1 and ADUL T2 concatenated.
BIRTH	9672	97	Records concerning a very wide number of readings and factors recorded at various stages during pregnancy . Most attributes are binary;and 70 of the attributes are very sparse;with over 95% of the values being FALSE.
SYNTH	30K--500K	24	Synthetic datasets of entirely binary attributes generated using bayes net.
CENSUS1	143,000	13	A larger (different) census dataset.
CENSUS2	143,000	16	An even larger census dataset with higher arity attributes.

AD-tree Version 3.0: Leaflists

If a node matches fewer than R_{min} nodes, don't build a subtree. Just record a list of pointers into the dataset.



Row	a_1	a_2	a_3
1	1	1	1
2	1	1	2
3	1	2	1
4	2	2	2
5	2	1	1
6	2	1	2
7	2	2	1
8	3	2	2
9	3	1	1

$R_{min} = 4$

...this can mean big memory savings ...

AD-tree size

Dataset	M	R	Nodes	#Mb	Build Time
CENSUS1	13	142521	24007	1.5	17
CENSUS2	15	142521	209577	13.2	32
ADUL T1	15	15060	58200	7.0	6
ADUL T2	15	30162	94900	10.9	10
ADUL T3	15	45222	162900	15.5	15
BIRTH	97	9672	87400	7.9	14
SYN30K	24	30000	34100	2.8	8
SYN60K	24	60000	59300	4.9	17
SYN125K	24	125000	95400	7.9	36
SYN250K	24	250000	150000	12.4	73
SYN500K	24	500000	219000	18.2	150

Bayes Net Results

Dataset	M	R	AD-tree Time	Regular Time	Speedup
CENSUS1	13	142521	162	48300	298.1
CENSUS2	15	142521	783	27000	34.5
ADUL T1	15	15060	150	2850	19
ADUL T2	15	30162	226	5160	22.8
ADUL T3	15	45222	220	7140	32.5
BIRTH	97	9672	23	2820	122.6
SYN30K	24	30000	32	10410	325.3
SYN60K	24	60000	32	18360	573.8
SYN125K	24	125000	34	42840	1260.0
SYN250K	24	250000	35	88830	2538.0
SYN500K	24	500000	36	5158	4298.3

Effect of R_{min}

	ADUL T2				BIRTH			
R_{min}	#Mb	#nodes	Build	Search	#Mb	#nodes	Build	Search
1	10.87	94872	13	225				
2	8.46	86680	11	223				
4	6.30	75011	9	223	27.60	245722	32	25
8	4.62	62095	7	224	14.90	152409	21	22
16	3.37	49000	6	232	7.95	87415	15	23
32	2.55	37790	5	245	4.30	46777	11	26
64	1.98	27726	5	274	2.42	23474	8	31
128	1.59	18903	3	331	1.48	11554	7	38
256	1.38	12539	3	420	0.95	5150	5	61
512	1.21	7336	3	586	0.65	2237	4	100
1024	1.05	3928	3	881	0.45	887	3	170
2048	0.95	1890	1	1231	0.28	246	1	270
4096	0.86	780	1	1827	0.28	201	2	303

Feature Selection

Given a contingency table for Area, HoursPerWeek and Sex, we can see how well Area and Sex predicts HoursPerWeek.

```
adtree> matrix
   sex        area  |     v0:20-     v1:20+     v2:40+
--------------------------------------------------------
Female      Abroad  |        237         27         34
Female     Midwest  |       1329        261        999
Female   Northeast  |        891        174        623
Female       South  |       1996        343       1180
Female     Unknown  |      37326       3955      21863
Female        West  |       1697        301        940
  Male      Abroad  |        188         32         95
  Male     Midwest  |       1077        249       1214
  Male   Northeast  |        625        149        717
  Male       South  |       1385        330       1537
  Male     Unknown  |      28645       3324      25837
  Male        West  |       1288        305       1348

Entropy(hours-per-week) = 1.26995
IG( hours-per-week | sex area ) = 0.011772
```

...Not very well!

Feature Selection

A better predictor of HoursPerWeek:

occupation	v0:20-	v1:20+	v2:40+
Adm-support-clerical	1054	1124	8570
Armed-Forces	11	6	9
Executive-and-managerial	358	504	7902
Farming-forestry-and-fishing	403	324	1598
Handlers-equip-cleaners-etc	624	521	1861
Machine-operators-assmblrs	508	535	3540
Other-service	1692	1448	5662
Precision-production-craft	596	905	6118
Private-household-services	171	103	298
Professional-specialty	556	1084	7916
Protective-services	83	90	1008
Sales	1065	913	6427
Technicians-and-related	122	192	1860
Transportation-and-moving	243	421	2275
Unknown	69198	1280	1343

Entropy(hours-per-week) = 1.26995
IG(hours-per-week | occupation) = 0.66545

Feature Selection

- Best 1-feature predictor of hours-per-week: occupation (IG 0.66)
- 2 Features: Occupation Parents (IG 0.69)
- 3 Features: Occupation Parents Age (IG 0.708)
- 4 Features: Occupation Parents Age MaritalStatus (IG 0.716)
- 5 Features: Occupation Age MaritalStatus Area WorkClass (IG 0.733)

Finding the best set of N-features can be achieved by search over all attribute sets of size N+1 that also include the desired feature.

(Heuristic Greedy versions are more common). Overfit warning.

Feature Selection Results

ADUL T2

Number Attributes	Number of Attribute Sets	AD-tree Time	Regular Time	Speedup
1	14	.000071	.048	675.0
2	91	.00054	.067	124.0
3	364	.0025	.088	34.9
4	1001	.0083	.11	13.4
5	2002	.023	.14	6.0

BIRTH

Number Attributes	Number of Attribute Sets	AD-tree Time	Regular Time	Speedup
1	96	.000018	.015	841
2	4,560	.000042	.021	509
3	142,880	.000093	.028	298
4	3,321,960	.00019		
5	61,124,064	.00033		

Output from Rule Finder

(Applied to ADULT2 dataset)

Rule 0: score = 0.965 (218/226), workclass = Private, education-num = above12, marital-status = Married-civ-spouse, capital-loss = above1600

Rule 1: score = 0.961 (199/207), workclass = Private, education-num = above12, relationship = Husband, capital-loss = above1600

Rule 2: score = 0.955 (193/202), age = below45, education-num = above12, relationship = Husband, capital-loss = above1600

Rule 3: score = 0.950 (302/318), education-num = above12, relationship = Husband, capital-loss = above1600, native-country = United-States

Rule 4: score = 0.948 (201/212), education-num = above12, relationship = Husband, capital-loss = above1600, hours-per-week = above40

Rule Learning Results

ADUL T2

Number Attributes	Number of Rules	AD-tree Time	Regular Time	Speedup
1	116	.000019	.0056	295.0
2	4,251	.000019	.0014	75.3
3	56,775	.000024	.00058	23.8
4	378,984	.000031	.00030	9.8
5	1,505,763	.000042	.00019	4.7

BIRTH

Number Attributes	Number of Rules	AD-tree Time	Regular Time	Speedup
1	194	.000025	.0072	286
2	17,738	.000021	.0055	259
3	987,134	.000022	.0040	186
4	37,824,734	.000024	.0030	127
5	1,077,672,055	.000026		

Drawbacks

- Don’t forget. This is still an **exponential time** and memory method (exponential in number of attributes).
- AD-Tree gets **huge if attributes are uncorrelated**. More irritating than it might appear.
- Doesn’t do **real-valued attributes**.
- Our experiments have compared against brutal counting approaches. Note that many brute counting algorithms can be sped up by **algorithm-specific tricks**.
- If leaf-lists are used, must keep the **dataset in main memory**.
- **Cannot cheaply add** records to a pre-existing AD-tree.

Related Work

- Decision Trees
- Kd-trees
- R-trees *[Guttman 1984]* (Proceedings of SIGMOD 84)
- Frequent Sets *[Agrawal, Mannila, Srikant, Toivonen and Verkamo , 1996]* (KDD96). Counting using Frequent sets: *[Mannila and Toivonen, 1996]* (KDD96)
- Datacube *[Harinarayan, Rajaraman and Ullman, SIGMOD1996]*
- Sampling
- "Remember singleton and doubleton contingency tables. Compute the rest."

ADT ree improvements to come...

- Version 4.0: Clever orderings of attributes .
- On-disk version of AD-trees
- Real-valued attributes ...

  All Real? :-)

  One or two real? :-)

  Honest mix of Real/Symbolic? ;-\
- Lazy Adtrees
- Shallow Adtrees
- Some standardization of whether it’ s “AD-Tree” “ ADTree” “ AdTree” “ad-tree” or “Adtree”

Future Exploitations of AD-Trees

- Faster Rule learners *[Brigham Anderson]*
- Schemata-search feature selectors
- Practical, much faster, Bayes Net learners
- Exhaustive Decision Tree search?
- Memory-Based Probabilistic Inference
- Logistic Regression
- Text Classification
- Propositional Feature Construction
- Efficient Bitstring search (GA style search)
- Optimal Variable Assignment FIT™

Neural Nets, Nearest Neighbor: Harder to see how to use AD-Trees.

For more information

www .cs.cmu.edu/~A UT ON

AD-tree source code (ANSI C) will be available from November 30th. (Above address)