Tree Induction vs. Logistic Regression
for Learning Rankings based on Likelihood of Class Membership

Claudia Perlich
New York University

Work with: Foster Provost & Jeff Simonoff
NIPS-2002

(*paper to appear in JMLR*)

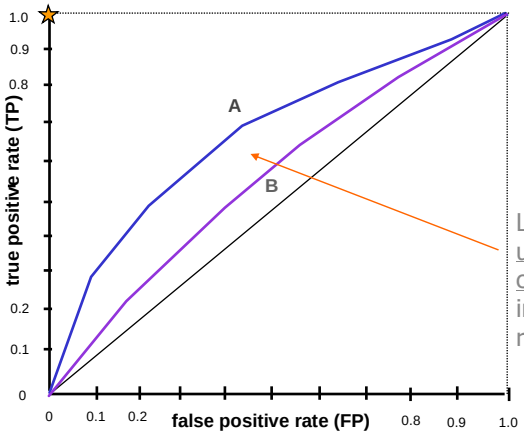
How do common learning algorithms compare for building models for ranking?
(based on likelihood of class membership)

- Logistic regression is an obvious candidate
- Tree induction?
  - attractive: ease of application, fast, comprehensible (arguably), "infinite capacity"
  - criticized for producing poor scores
- Tree induction for 0/1 loss (accuracy)?
  - Lim, Loh & Shih (MLJ 2000) compared on 32 data sets
    - Logistic regression "beats" C4.5 (7% lower average error rate)
    - Logistic regression is 2nd "best" algorithm
      - and best one is impracticable
    - C4.5 did not perform particularly well
      - 17th best
- We'll "fix" tree induction a bit, then use two analytical tools to compare tree induction and logistic regression.

Analytical Tool #1: ROC Curves and AUR

"Receiver Operating Characteristic" analysis (from signal detection theory)

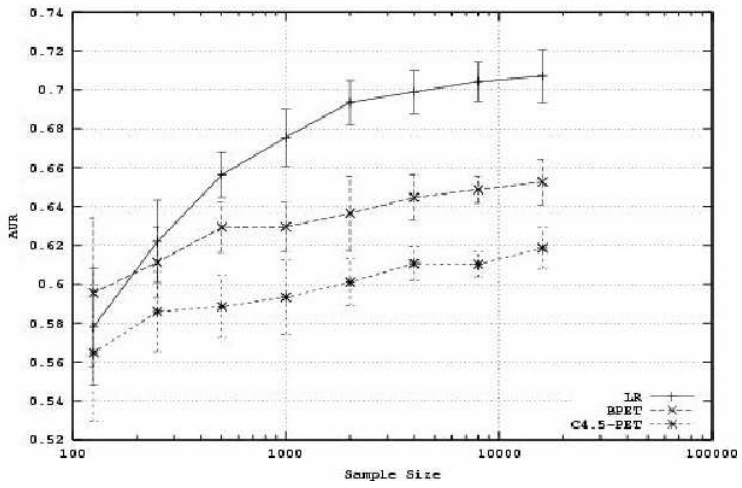
ranking models produce a range of possible (FP,TP) tradeoffs



Larger area under the ROC curve (AUR) implies better ranking model

- Separate classifier performance from costs and target class distributions
- AUR equivalent to Wilcoxon-Mann-Whitney statistic & (essentially) the Gini coefficient (Hand, 1997)

Analytical tool #2: learning curves



Tree induction vs. Logistic Regression

- Combine learning-curve analysis and AUR
- Massive experimental study: 36 data sets
- Use large data sets
  - mean training size: 60,000 (median = 12,800)
  - (LLS MLJ-2000 study: mean = 900)
- Look at class-based ranking (& classification)
- Examine effect(s) of data-set size
  - Example questions:
    - can trees be competitive for class-based ranking?
    - is LR better for smaller training sets?
    - are different algorithms better on different types of data?

Tree Induction & Logistic Regression

- Tree induction
  - C4.5
  - C4.5-PET (Probability Estimation Tree - No pruning, Laplace correction)
  - BPET (Averaged-bagging of C4.5-PETs)

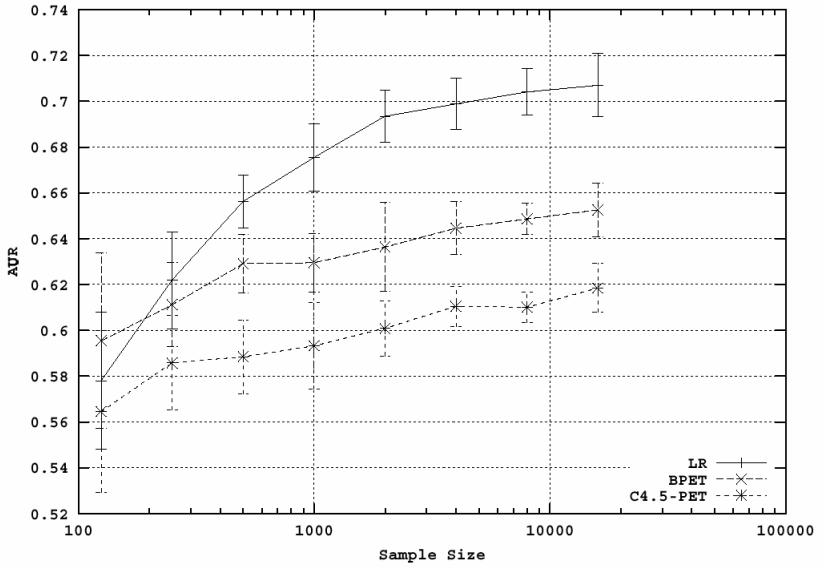
  → C4.5-PET and BPET generally improve performance for ranking, [cf. (Provost, Fawcett, Kohavi, ICML-98); (Bauer & Kohavi, MLJ 1999); (Provost & Domingos, MLJ to appear)]
- Logistic regression
  - as implemented in SAS (also tested R and Splus versions)
  - Model selection (various methods)
  - Ridge regression
  - Bagging

  → Variants generally help very little or hurt on larger sample sizes, so we'll consider only regular logistic regression

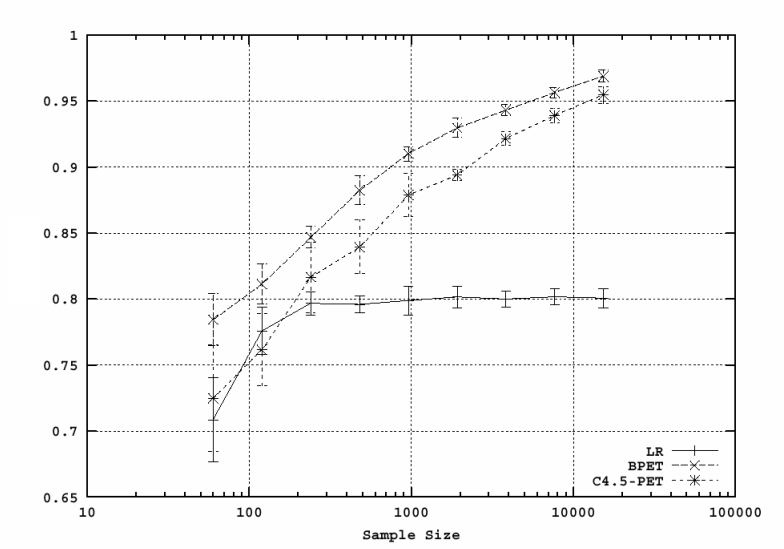
Tree Induction vs. Logistic Regression for producing ranking models

- Result categorization (36 data sets)
  - Learning curves indistinguishable (9 cases)
  - Logistic Regression ultimately better (10 cases)
  - Tree Induction ultimately better (17 cases)

Logistic Regression Dominates

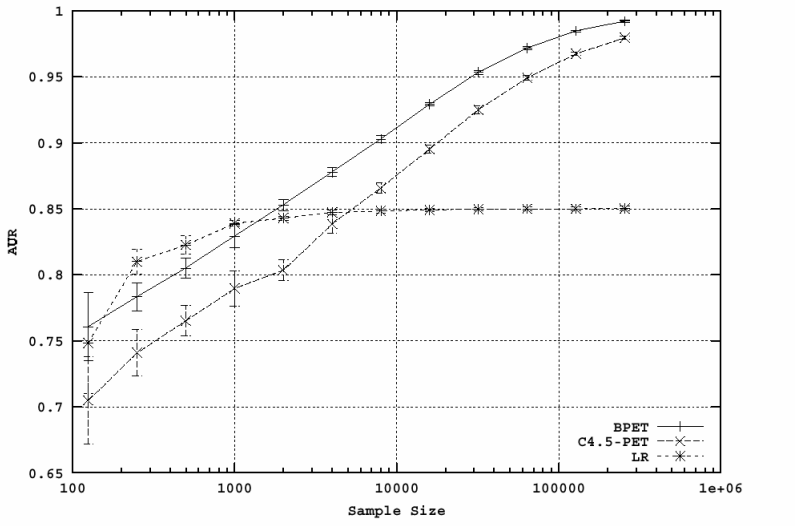


Tree Induction Dominates



9

Tree Induction Crosses



Claudia Perlich and Foster Provost

10

TI-LR

High separability
16-2

Low separability
1-8

Data set	Winner AUR	Winner Acc	Max AUR	Result
Nurse	none	none	1	Indistinguishable
Mushrooms	none	none	1	Indistinguishable
Optdigit	none	none	0.99	Indistinguishable
Letter–V	C4	C4	0.99	C4 dominates
Letter–A	C4	C4	0.99	C4 crosses
Intrusion	C4	C4	0.99	C4 dominates
DNA	C4	C4	0.99	C4 dominates
Covertype	C4	C4	0.99	C4 crosses
Telecom	C4	C4	0.98	C4 dominates
Pendigit	C4	C4	0.98	C4 dominates
Pageblock	C4	C4	0.98	C4 crosses
CarEval	none	C4	0.98	C4 crosses
Spam	C4	C4	0.97	C4 dominates
Chess	C4	C4	0.95	C4 dominates
CalHous	C4	C4	0.95	C4 crosses
Ailerons	none	C4	0.95	C4 crosses
Firm	LR	LR	0.93	LR crosses
Credit	C4	C4	0.93	C4 dominates
Adult	LR	C4	0.9	Mixed
Connects	C4	none	0.87	C4 crosses
Move	C4	C4	0.85	C4 dominates
Downsize	C4	C4	0.85	C4 crosses
Coding	C4	C4	0.85	C4 crosses
German	LR	LR	0.8	LR dominates
Diabetes	LR	LR	0.8	LR dominates
Bookbinder	LR	LR	0.8	LR crosses
Bacteria	none	C4	0.79	C4 crosses
Yeast	*none*	*none*	*0.78*	*Indistinguishable*
Patent	C4	C4	0.75	C4 crosses
Contra	*none*	*none*	*0.73*	*Indistinguishable*
IntShop	LR	LR	0.7	LR crosses
IntCensor	LR	LR	0.7	LR dominates
Insurance	*none*	*none*	*0.7*	*Indistinguishable*
IntPriv	LR	none	0.66	LR crosses
Mailing	LR	none	0.61	LR dominates
Abalone	LR	LR	0.56	LR dominates

11

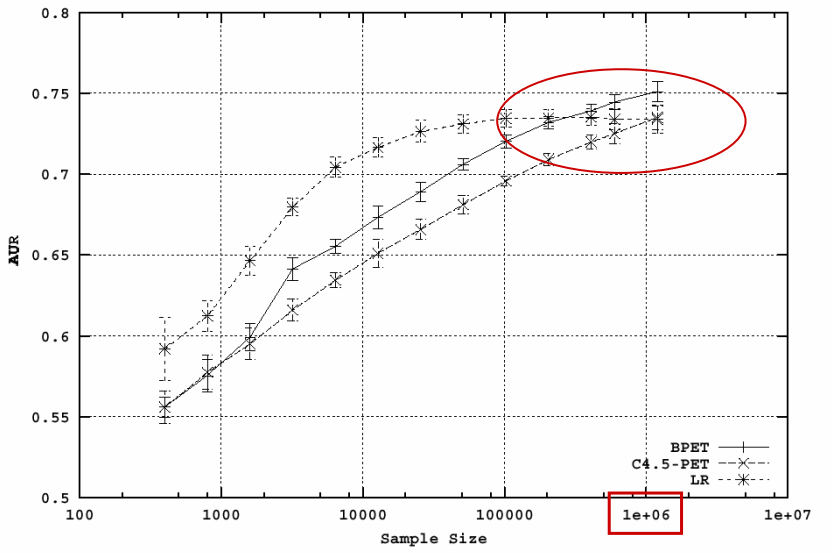
Implications I

- *Logistic regression does not generally outperform tree induction*
  - contrary to the results of Lim, Loh & Shih (MLJ 2000)
  - logistic regression often is better for smaller training sets
  - tree induction often is better for larger training sets
- *Tree induction is remarkably effective at producing class-based rankings*
  - contrary to conventional wisdom
- *Learning from large data sets is justified*
  - tree-induction learning curves keep increasing

Claudia Perlich and Foster Provost

12

Massive training sets?



13

Implications II

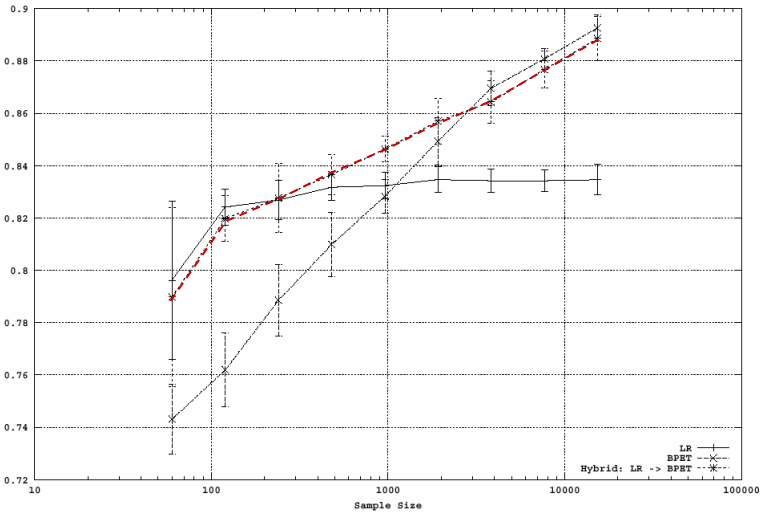
- *Must exercise care when drawing conclusions about algorithm superiority for a particular application*
  - learning curves cross
  - conclusions about superiority for an application must be based on an analysis of the learning curves

AUR: 23-11-2 vs. 10-9-17

Acc: 22-13-1 vs. 8-14-14

- *Tree induction and logistic regression are preferable (ultimately) for different kinds of data*
  - Tree Induction for high-separability data
  - Linear Regression for low-separability data

Food for thought: A hybrid algorithm?



References

[NB: All but the first two can be obtained from http://pages.stern.nyu.edu/~fprovost]

- Bauer, E. and R. Kohavi (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning* 36, 109-139.
- Lim, T., W. Loh, Y. Shih (2000). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. *Machine Learning*, 40(3), 203-228.
- Perlich, C., Provost F., and Simonoff, J. S. (2001). Tree induction vs. logistic regression: A learning-curve analysis. To appear in *Journal of Machine Learning Research*.
- Provost, F. and P. Domingos. Tree Induction for Probability-based Rankings. To appear in *Machine Learning*.
- Provost, F., & Fawcett, T (2001). Robust classification for imprecise environments. *Machine Learning,* 42, 203-231.
- Provost, F., T. Fawcett, and R. Kohavi (1998). The Case Against Accuracy Estimation for Comparing Classifiers. ICML-98.
- Saar-Tsechansky, M. and F. Provost (2002). Active Sampling for Class Probability Estimation and Ranking. To appear in *Machine Learning*.

Tree induction: comparison

- Accuracy (win-tie-loss)
  - C4.5 beats PET: 10-25-1, but improvements small
  - BPET beats C4.5: 10-21-5, some improvements substantial
- Ranking/probability estimation
  - PET beats C4.5: 22-12-2
  - BPET beats C4.5: 24-12-0
  - BPET beats PET: 15-19-2

→ C4.5-PET and BPET generally improve performance for ranking, [cf. (Provost, Fawcett, Kohavi, ICML-98); (Bauer & Kohavi, MLJ 1999); (Provost & Domingos, MLJ to appear)]

- Conclusion: we'll use C4.5-PET and BPET for comparison with logistic regression

Example: Spam

Logistic regression: comparison

- Logistic regression is remarkably robust
- Model selection and Ridge regression
  - help for small data sets
  - no difference after a few thousand examples
- Bagging
  - systematically detrimental

→ Variants generally help very little or hurt on larger sample sizes, so we'll consider only regular logistic regression

- Conclusion: we'll use standard logistic regression for comparison with tree induction

Example: California Housing