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.
Claudia Perlich and Foster Provost

**Analytical Tool #1: ROC Curves and AUR**

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

```
A
B
```

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

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

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

```
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
```

true positive rate (TP)
false positive rate (FP)

---

**Analytical tool #2: learning curves**

```
Sample Size
1.02
1.04
1.06
1.08
1.10
```

```
AUC
```

Claudia Perlich and Foster Provost
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

Tree Induction Crosses
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
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]


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

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: Spam

Example: California Housing