

Combining Boosting with Trees for the KDD Cup 2009

`dmclab@i6.informatik.rwth-aachen.de`

June 28, 2009

**Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany**

Outline

- ▶ **Task Description**

- ▶ **Preprocessing**
 - Missing Values**
 - Feature Generation & Selection**

- ▶ **Classification**
 - Boosted Decision Stumps**
 - Logistic Model Tree**

- ▶ **Combinations**
 - AUC based optimizations**

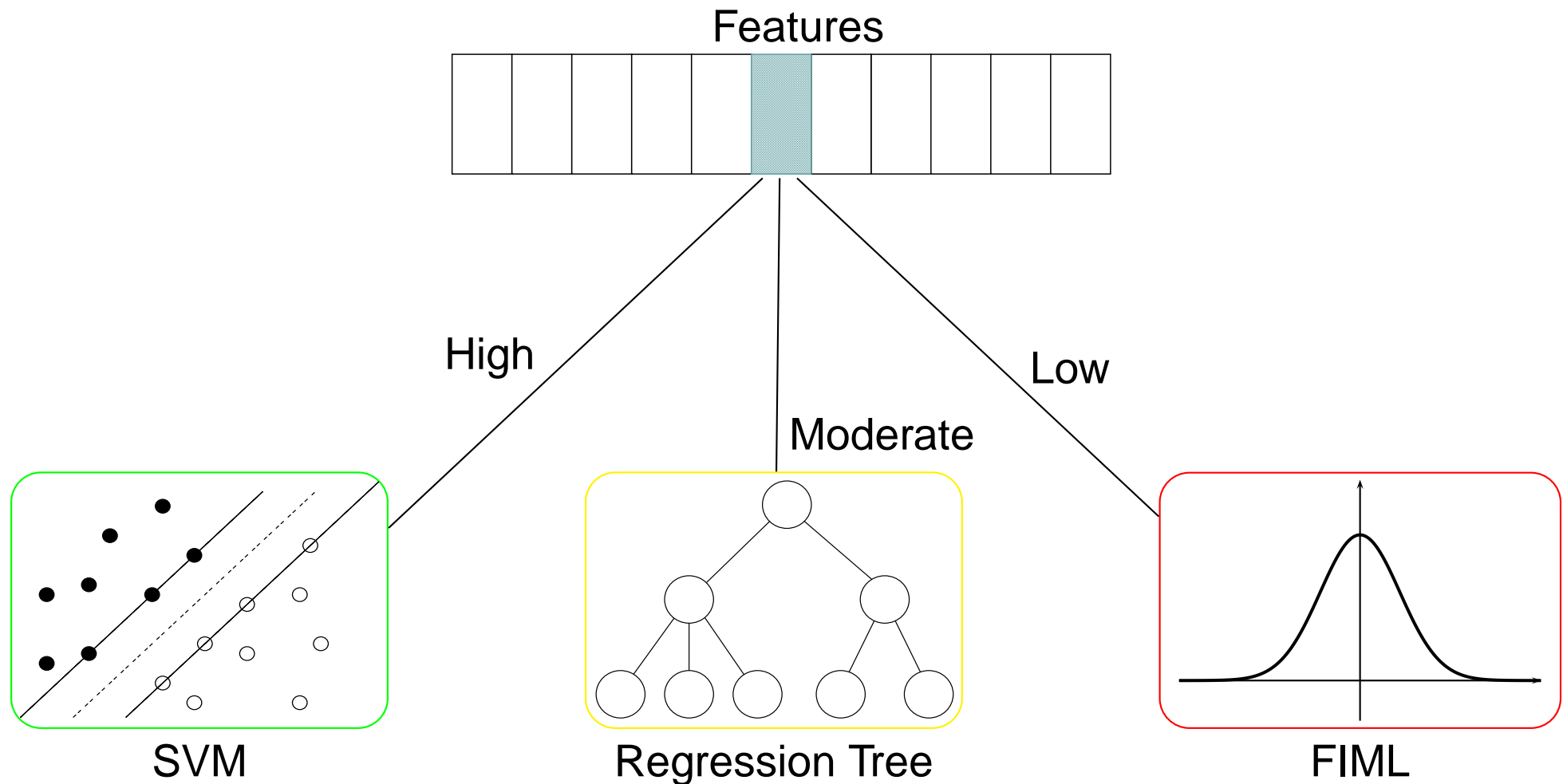
- ▶ **Conclusions**

The Small Challenge

- ▶ **RWTH Aachen Data Mining Lab**
Organized since 2004
This year first participation in the KDD Cup with eight students
- ▶ **Slow track with small data set**
230 features (190 numerical and 40 categorical)
32 days duration
- ▶ **Best submission without unscrambling**
Ranked 35th in final evaluation

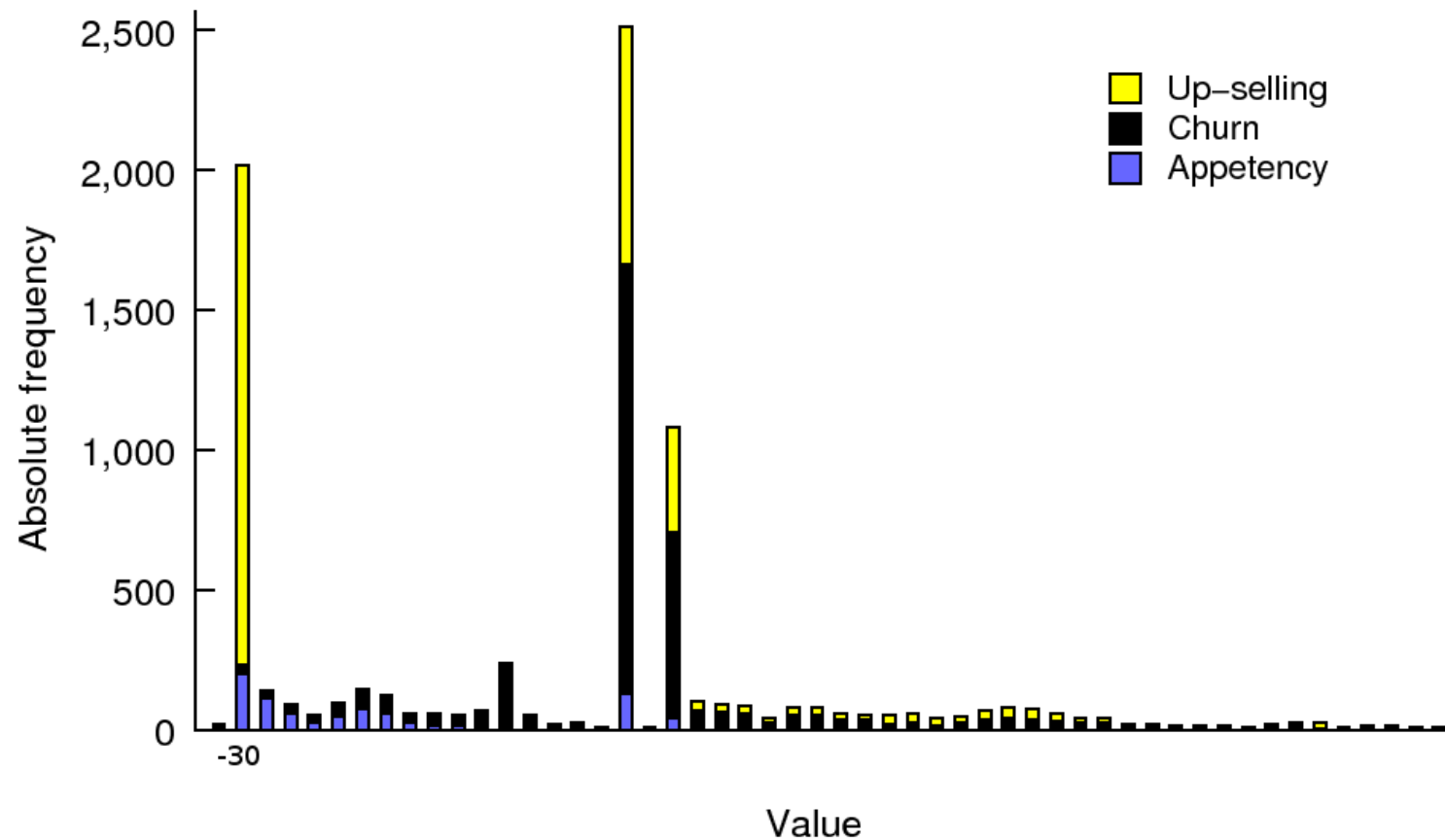
Preprocessing: Missing Values

- ▶ **Missing Value Ratio (MVR) for feature f :**
average number of samples per class and feature where f is not missing



Preprocessing: Features

► Generation of binary features

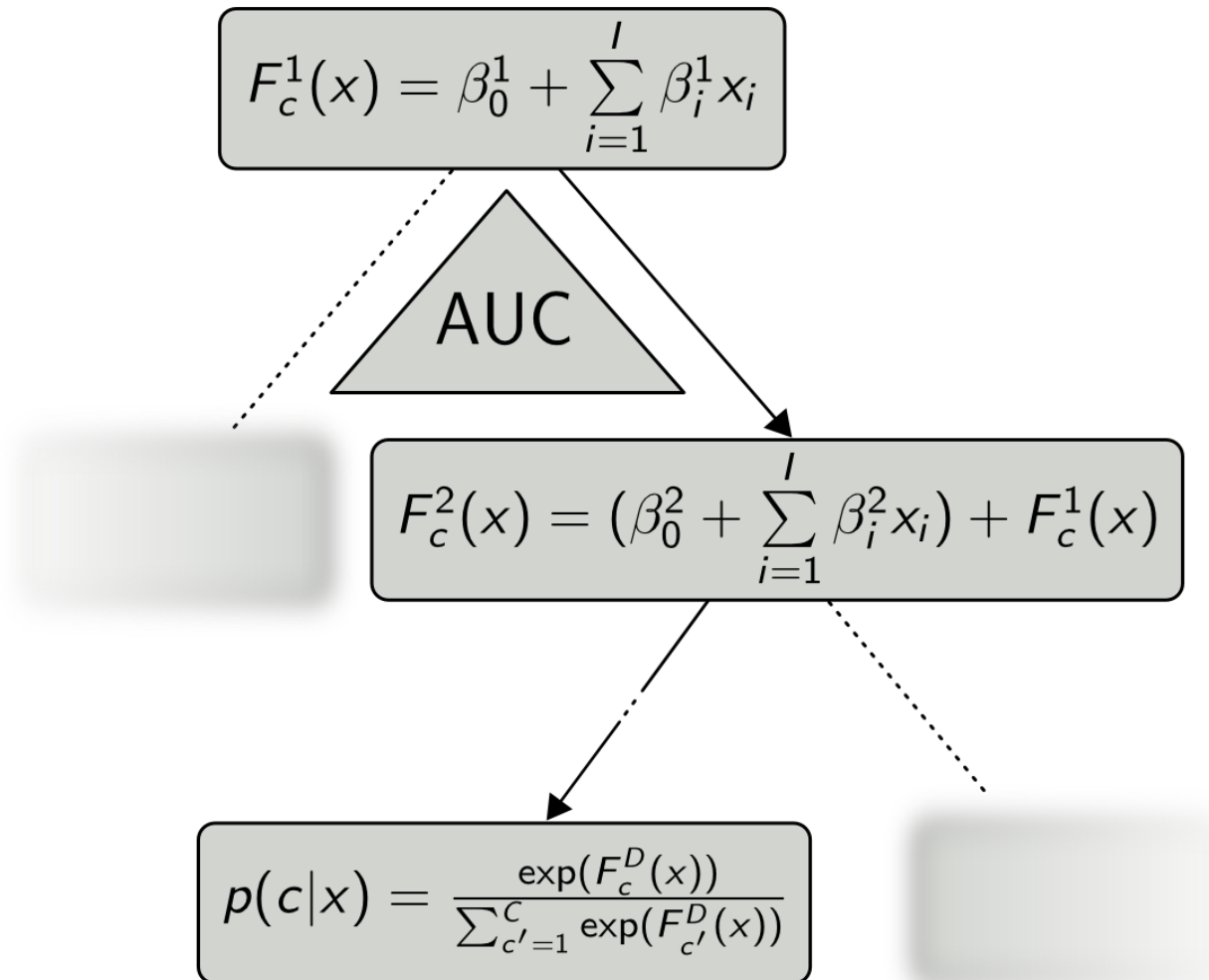


► Feature Selection: Ranking based on information gain and likelihood-ratio

Boosted Decision Stumps

- ▶ **AdaBoost with one-level decision trees as weak learners**
Implemented in Boostexter by Schapire and Singer [2000]
- ▶ **Linear complexity in training: $O(CN)$**
 C : number of classes
 N : number of training instances
- ▶ **Best performance as single classifier**

Logistic Model Tree

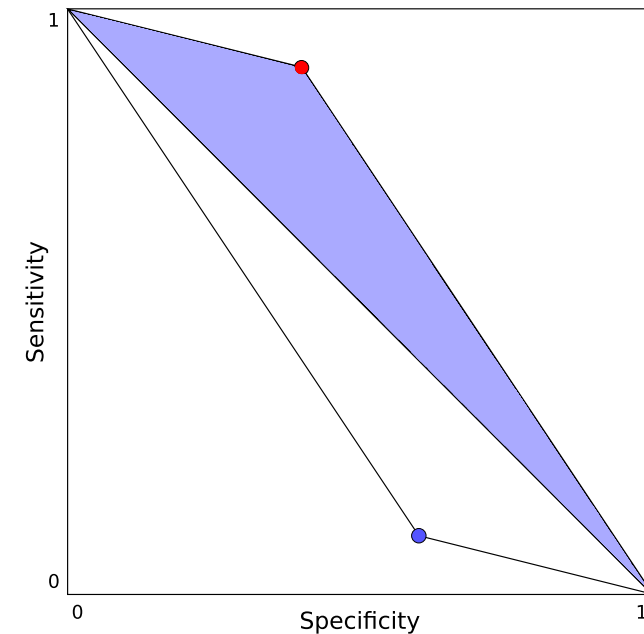
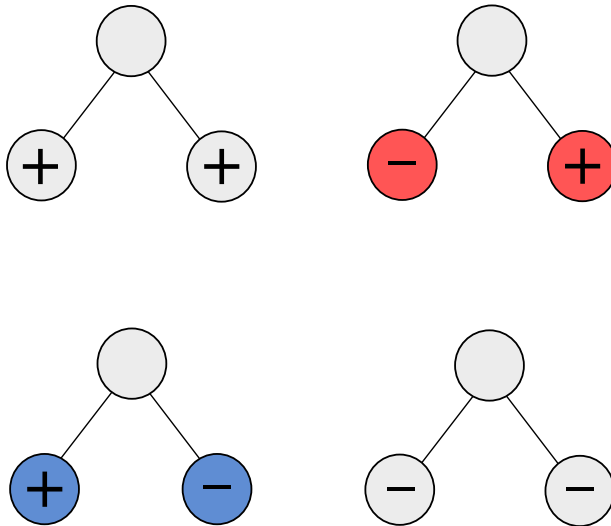


F_c^d : linear regression of observation vector x in node d for class c

β_i^d : regression coefficient for the i th component of x in node d

AUC Split Criterion

- ▶ Introduced by Ferri et al. [2003] for trees with two classes (+) and (-)
- ▶ Each labeling of the leaves corresponds to one point in the ROC space



- ▶ **Local Positive Accuracy:** $\frac{N_l^+}{N_l^+ + N_l^-}$

N_l^c : Number of training samples in leaf l assigned to class c

- ▶ **Select split point resulting in largest AUC**

Combinations

- ▶ **Stacking**

 - Predictions of boosted decision stumps as features for Logistic Model Tree**

- ▶ **Linear combinations of predictions, optimizing on the AUC**

 - Weighted scores**

 - Weighted voting**

Results

► AUC score of single classifiers on cross-validation

Classifier	Appetency	Churn	Up-selling	Score
Boosted decision stumps	0.8172	0.7254	0.8488	0.7971
Logistic Model Tree	0.8176	0.7161	0.8450	0.7929
Multilayer perceptron	0.8175	0.7049	0.7741	0.7655

► Combination of LMT, MLP and boosted decision stumps

Combination method	Appetency	Churn	Up-selling	Score
Weighted scores	0.8256	0.7306	0.8493	0.8018
Weighted votes	0.8225	0.7331	0.8515	0.8023

Conclusions

- ▶ **Best performance: Boosted decision stumps and Logistic Model Tree**
- ▶ **Combinations and stacking**
- ▶ **AUC-optimized combinations**
- ▶ **Results in KDD Cup**

Rank	Method	Appetency	Churn	Up-selling	Score
35	LMT + AUCsplit	0.8268	0.7359	0.8615	0.8080
36	Weighted votes	0.8204	0.7398	0.8621	0.8074

Thank you for your attention

Patrick Doetsch

patrick.doetsch@rwth-aachen.de

<http://www-i6.informatik.rwth-aachen.de/>

References

- C. Ferri, P. A. Flach, and J. Hernandez-Orallo. Improving the auc of probabilistic estimation trees. In *Proceedings of the 14th European Conference on Machine Learning*, pages 121–132. Springer, 2003. 8
- R. E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000. 6