

An Ensemble of Three Classifiers for KDD Cup 2009: Expanded Linear Model, Heterogeneous Boosting, and Selective Naive Bayes

Team: *CSIE Department, National Taiwan University*

Members: Hung-Yi Lo, Kai-Wei Chang, Shang-Tse Chen, Tsung-Hsien Chiang, Chun-Sung Ferng, Cho-Jui Hsieh, Yi-Kuang Ko, **Tsung-Ting Kuo**, Hung-Che Lai, Ken-Yi Lin, Chia-Hsuan Wang, Hsiang-Fu Yu, Chih-Jen Lin, Hsuan-Tien Lin, Shou-de Lin



DEPT. OF COMPUTER SCIENCE AND INFORMATION ENGINEERING
GRADUATE INST. OF NETWORKING AND MULTIMEDIA

2009/6/26



Observations on Training

- Positive labels of 3 tasks are exclusive
 - Transform to 4-class classification (C / A / U / Null)

	Label of churn	Label of appetency	Label of up-selling
class 1	1	0	0
class 2	0	1	0
class 3	0	0	1
class 4	0	0	0

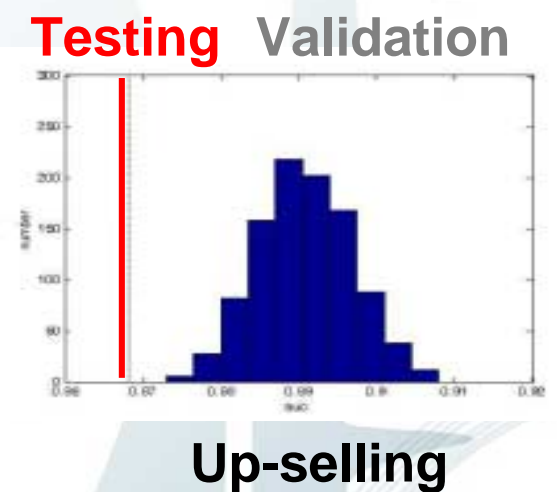
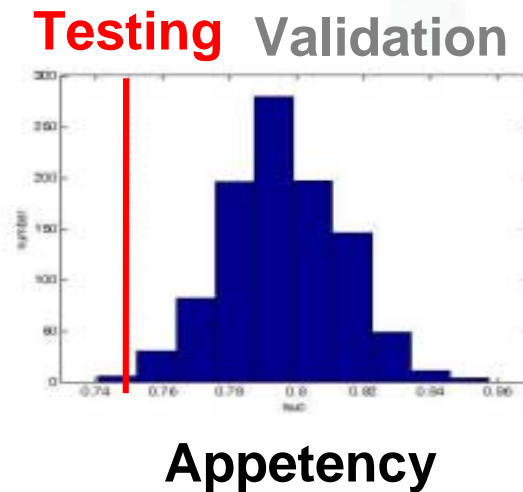
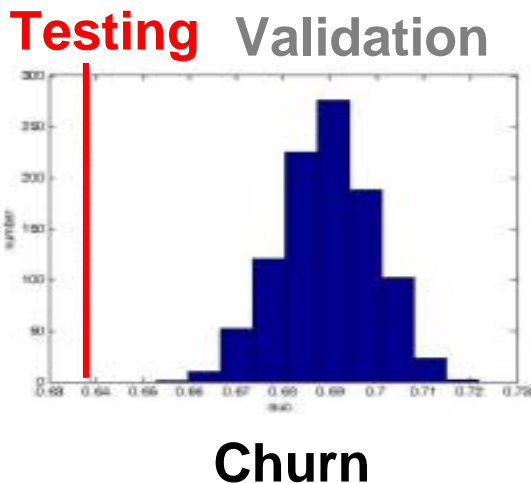
2009/6/26

2



Observations on 10% Testing

- Cross-validation (CV) varies from 10% testing (online feedback) significantly for certain classifiers
 - Use CV instead of 10% results to tuning up parameters



2009/6/26

3



Challenges

1. Noisy

- Redundant or irrelevant features: **feature selection**
- Significant amount of **missing values**

2. Heterogeneous

- Number of distinct **numerical** values: 1 to ~50,000
- Number of distinct **categorical** values: 1 to ~15,000

→ **Decision of classifiers and pre-processing methods !**



System Overview

Pre-processing

Feature Selection

Missing Values

Numerical Features

Categorical Features

Classification

Maximum Entropy

Heterogeneous AdaBoost

Selective Naïve Bayes

Post-processing

Score Adjustment

Score Ensemble

2009/6/26

5



Maximum Entropy

- Transform to joint multi-class classification
 - Maximum entropy model → probability estimation
- Feature selection
 - L1-regularized solver

$$\sum_{y=1}^k \sum_{t=1}^n |w_{yt}| + C \log \sum_{i=1}^l \frac{e^{w_{y_i}^T \mathbf{x}_i}}{\sum_{y=1}^k e^{w_y^T \mathbf{x}_i}}$$

L1
Regularization

Maximum
Entropy

x = example
 y = label
 w = model
 k = # of classes
 n = # of iterations
 l = # of examples
 C = penalty parameter

2009/6/26

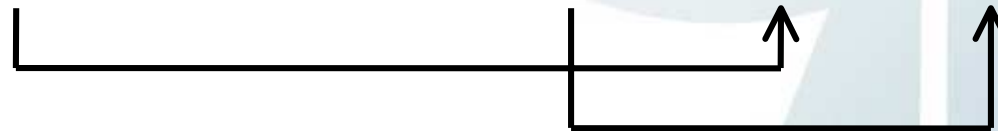
6



Max-Ent: Missing Values

- Fill missing values with zeroes or “missing”
- Add a binary feature to indicate “missing”

ID	Label	Num1	Num2	Cat1	Cat2	Mis1	Mis2
1	+1	10	0.2	A	D	0	0
2	-1	2	0.4	A	Miss	0	1
3	-1	100	0.5	B	Miss	0	1
4	+1	0	0.3	C	E	1	0
5	1	20	0.1	B	Miss	0	1



2009/6/26

7



Max-Ent: Numerical Feature

- Log scaling
- Linear scaling to $[0, 1]$

ID	Num1	Num2
1	10	0.2
2	2	0.4
3	100	0.5
4	0	0.3
5	20	0.1



ID	Log1	Log2	Lin1	Lin2
1	1.000	-0.699	0.100	0.400
2	0.301	-0.398	0.020	0.800
3	2.000	-0.301	1.000	1.000
4	0.000	-0.523	0.000	0.600
5	1.301	-1.000	0.200	0.200

2009/6/26

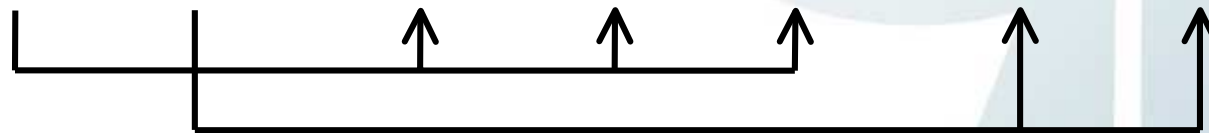
8



Max-Ent: Categorical Feature

- Add a binary feature for each category
- Also for numerical features with <5 distinct values

ID	Cat1	Cat2	A	B	C	D	E
1	A	D	1	0	0	1	0
2	A	Miss	1	0	0	0	0
3	B	Miss	0	1	0	0	0
4	C	E	0	0	1	0	1
5	B	Miss	0	1	0	0	0



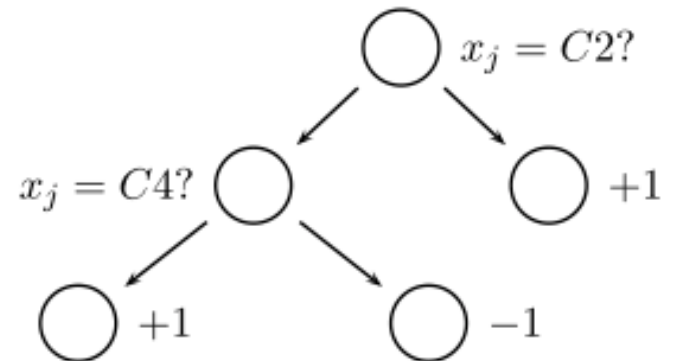
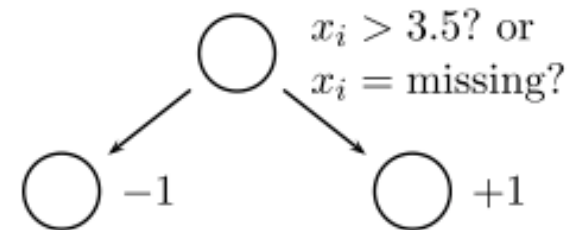
2009/6/26

9



Heterogeneous AdaBoost

- Feature selection
 - Inherent
- Missing value
 - Treated as a category
- Numerical feature
 - Numerical tree base learner
- Categorical feature
 - Categorical tree base learner
 - Height limitation for complexity regularization





Selective Naïve Bayes

- Feature selection
 - Heuristic search [Boullé, 2007]
- Missing value
 - No assumption required
- Numerical feature
 - Discretization [Hue and Boullé, 2007]
- Categorical feature
 - Grouping [Hue and Boullé, 2007]

2009/6/26

11



Score Adjustment

- Train a linear SVM to select one from the 4 classes
- For each classifier, generate features using
 - Scores of 3 classes
 - Entropy of the prediction scores
 - Ranks of 3 classes
- Use true label for training
- Output adjusted scores

2009/6/26

12



Score Ensemble

- Refer to the adjusted scores of **CV**
- Select **best 2** classifiers for each task
- **Average the rank** of scores from 2 classifiers
- Output the averaged rank as final result

2009/6/26

13



Results

- AUC Results
 - Train: CV
 - Test: 10% testing

Base Classifier	Churn		Appetency		Upselling		Score
	Train	Test	Train	Test	Train	Test	
Maximum Entropy	0.7326	0.7428	0.8669	0.8786	0.9001	0.8975	0.8396
Selective Naïve Bayes	0.7375	0.7428	0.8560	0.8529	0.8593	0.8564	0.8174
Heterogeneous AdaBoost	0.7350	0.7395	0.8660	0.8623	0.9030	0.9021	0.8347
Merged Model (w./o. PP)		0.7557		0.8671		0.9036	0.8421
Merged Model (w. PP)		0.7558		0.8789		0.9036	0.8461

Appetency improves most
with post-processing

2009/6/26

14



Other methods we have tried

- Rank logistic regression
 - Maximize AUC = maximize pair-wise ranking accuracy
 - Adopt pair-wise rank logistic regression
 - Not as good as other classifiers
- Tree-based composite classifier
 - Categorize examples using missing value pattern
 - Train a classifier for each of the 85 groups
 - Not significantly better than other classifiers

2009/6/26

15



Conclusions

- Identify **2 challenges** in data
 - Noisy → feature selection + missing value processing
 - Heterogeneous → numerical + categorical pre-processing
- Combine **3 classifiers** to solve the challenges
 - Maximum Entropy → convert data into numerical
 - Heterogeneous AdaBoost → combine heterogeneous info
 - Selective Naïve Bayes → discover probabilistic relations
- Observe **2 properties** from tasks
 - Training → model design and post-processing
 - 10% Testing → overfitting prevention using CV

Thanks !

2009/6/26

16



Reference

- [Boullé, 2007] M. Boullé. Compression-based averaging of selective naive bayes classifiers. *Journal of Machine Learning Research*, 8:1659–1685, 2007.
- [Hue and Boullé, 2007] C. Hue and M. Boullé. A new probabilistic approach in rank regression with optimal bayesian partitioning. *Journal of Machine Learning Research*, 8:2727–2754, 2007.