Bias Analysis in Text Classification for Highly Skewed Data

Lei Tang and Huan Liu

Data Mining & Machine Learning Lab
Department of Computer Science & Engineering
Arizon State University

ICDM 2005

Highly Skewed data in Text Categorization

Challenges

- Curse of Dimensionality
- Extremely Imbalanced (Major class : Minor Class > 67:1)

Existing Approaches

- Change the evaluation/optimization measure (ROC,F-measure)
- Sampling (over-sampling, under-sampling, change the threshold, cost-sensitive learning etc. [Kuba-Matw97, Bati-etal04])
- Feature Selection: Information Gain(IG), χ^2 [Yang-Pede97], Odds ratio [Mlad-Grob99] and Bi-normal separation(BNS) [Form03]

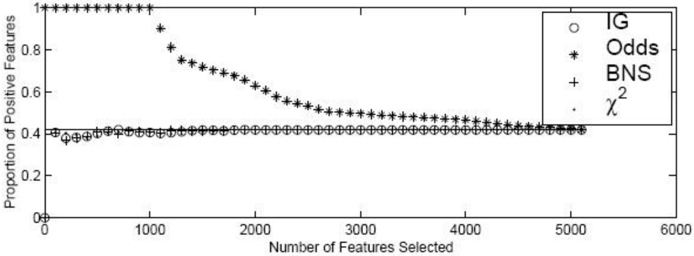
Introduction

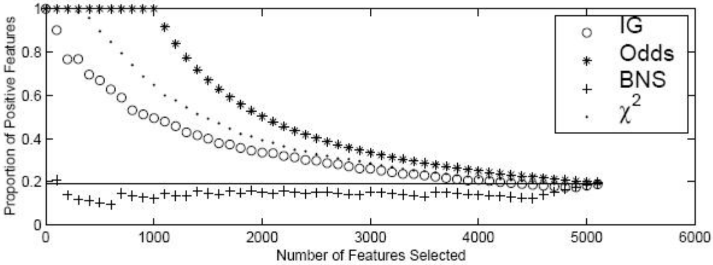
Here, we focus on binary skewed data with boolean attributes.
Two classes: Positive(Minor)/Negative(Major) class.

Different kinds of features

1. Positive features $P(f|+) > P(f|-)$
2. Negative features $P(f|+) < P(f|-)$
3. Neutral features $P(f|+) = P(f|-)$

Feature Selection Metric Bias



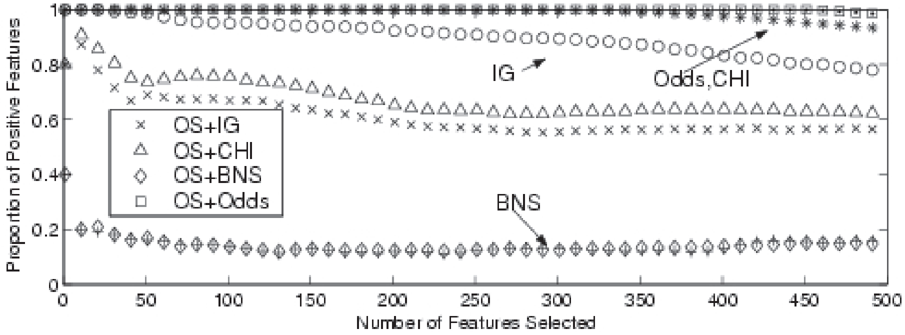


Various Biases

1. ***Feature Selection metric Bias***:
    - Biased metric: Odds ratio, Information Gain, χ^2 etc.
    - Unbiased metric(Bi-normal separation)
2. ***Class bias***: Classification favors major class (Use over-sampling to conquer the bias)
3. ***Classifier bias***:
    - Decision tree is embedded with feature selection and sensitive to sampling;
    - Naïve Bayes classifier is sensitive to sampling and feature selection;
    - SVM, moderately to both feature selection and sampling.

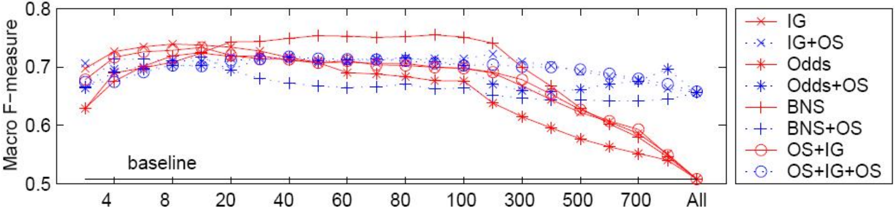
Investigate various biases in concert!!

Hypothesis I: Over-sampling vs. Metric Bias(1)



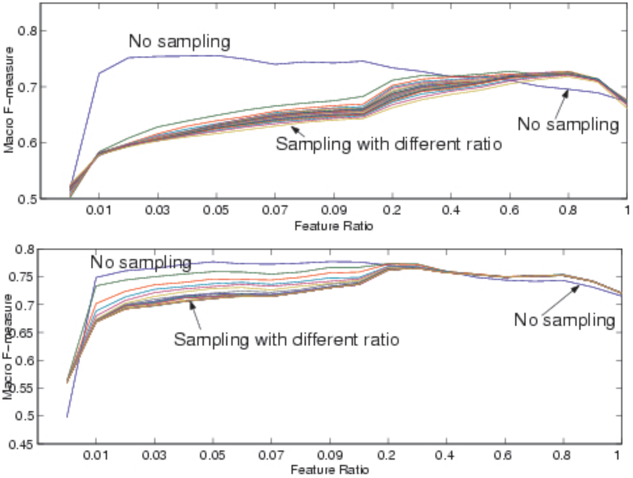
- over-sampling can make Information Gain to select more negative features.
- This, in certain degree, explains why over-sampling is always helpful for both pruned\unpruned trees.

Hypothesis I: Over-sampling vs. Metric Bias(2)



- Directly combine sampling with unbiased metric ↓.
- Usually, metric bias is more effective than over-sampling

Hypothesis II: Metric Bias vs. Sampling Ratio



Conclusions

1. Sampling before feature selection can cause selection of more negative features
2. It is more effective to select good features than to change the class distribution
3. The ratio between positive and negative features should be close to the class distribution
4. Biased feature selection metric plus sampling works fine.
5. Performance is insensitive to the sampling ratio if we do sampling after feature selection.

Bibliography

- Addressing the curse of imbalanced training sets: one-sided selection
  M. Kuba and S. Matwin , *ICML*, 1997.

- A Comparative Study on Feature Selection in Text Categorization
  Y. Yang and J. Pederson , *ICML*, 1997

- Feature Selection for Unbalanced Class Distribution and Naive Bayes
  D.Mlad and M.Grobelnik, *ICML*, 1999

- An extensive empirical study of feature selection metrics for text classification
  G.Forman, *J. Mach. Learn. Res.*, 2003

- A study of the behavior of several methods for balancing machine learning training data,
  G. Batista et al. *SIGKDD Explor. Newsl.*, 2004.