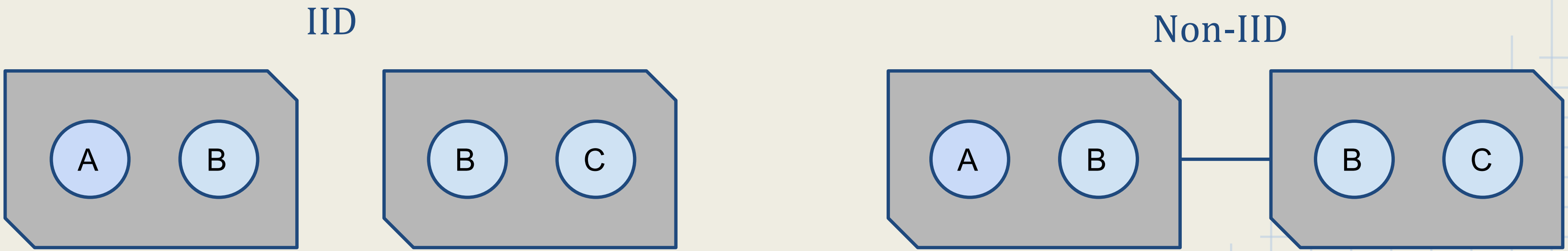
Zipf's Law and Relational Learning: An Investigation of a Surprising Correlation

Leo Behe and Zach Wheeler
Mentors: Dr. William Pottenger and Dr. Christie Nelson

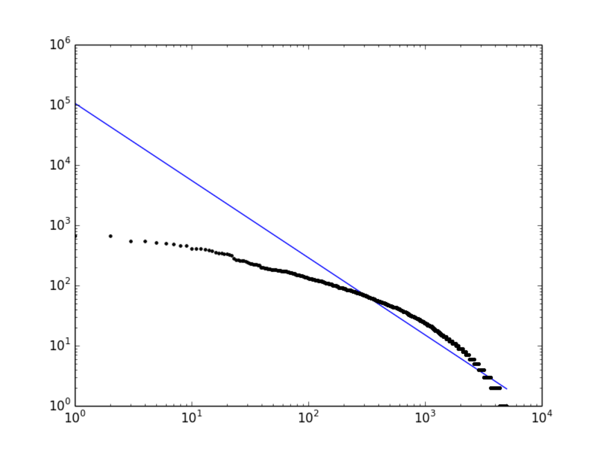
Overview: Relational Learning and IID

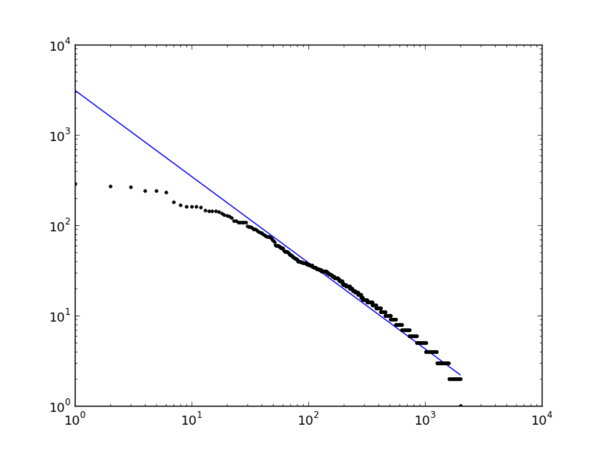
- Traditional machine learning algorithms make an "IID" assumption
- Relational learning does not make this assumption and leverages the latent relations between features



Zipf's Law

- Empirical law
  - Distribution of words in a natural language
  - Linear log-log plot of frequency to rank





Why Zipf?

- Correlation to relational learning?
  - Suggested by results from previous research
  - Our task: to investigate
- Importance
  - Correlation -> simple predictive test

Empirical Approach

- Compare NB (standard algo) to HONB (relational variant)
- Look at the performance of HONB in relation to the R-squared measure of the "Zipfian-ness" of the feature distribution
- Look for a correlation across different types of data, different n-gram sizes, different feature space sizes
    - N-grams are substrings of length N that are generated from text

Starting Hypotheses

- There is a correlation between the mean square error measurement of how Zipfian a dataset's feature distribution is and the classification performance of HONB.
- The effectiveness of HONB operating on a dataset composed of character n-gram features shows a statistically significant improvement as the character n-gram distribution approaches a Zipfian distribution.

Mathematical Approach

- HONB counts implicit paths of length k: optimal path length?
    - Involves counting SDRs of n given sets
- SDR: “System of Distinct Representatives”
    - Formula given for n=2,3,4
    - General case left open in 2011 paper
    - #P-Hard
    - Closed form might still be helpful theoretically

Mathematical Approach

Theorem:

$$\#SDR(A_1, A_2, ..., A_n) = \sum_{P \in \Pi_n} (-1)^{n-r(P)} \prod_{B \in P} (|B| - 1)! | \cap_{i \in B} A_i |$$

- Proof by Möbius inversion on the set partition lattice (order theory)

Literature Review

- *Ganiz, M.C.; George, C.; Pottenger, W.M.* - **Higher Order Naive Bayes: A Novel Non-IID Approach to Text Classification.** (2011, IEEE Transactions on Knowledge and Data Engineering)
    - First journal appearance of HONB
- *Dutta, S.; Sankaran, N.; Sankar K., P.; Jawahar, C.V.* - **Robust Recognition Of Degraded Documents Using Character N-Grams.** (2012, 10th IAPR International Workshop on Document Analysis Systems)
- *Unpublished research by Knopp, B.D.; Nelson, C.; Pottenger, W.M.* - **Modeling Microtext with Character n-grams and Higher Order Learning.**
- *Cavnar, W.B.; Trenkle, J.M.* - **N-Gram-Based Text Categorization.** (1994, 3rd Annual Symposium on Document Analysis and Information Retrieval)

Datasets we tested

- Ushahidi Haitian Earthquake Dataset
  - 2528 instances, 5 classes
  - Confirmed prior research in this area
- Twitter Sentiment Dataset
  - 1040 instances, 3 classes
- Washington Transportation Dataset
  - 1368 instances, 3 classes
  - Contributed by Dr. Williams
- California Transportation Dataset
  - 1221 instances, 3 classes
  - Contributed by Dr. Williams

Datasets we tested

- Worldwide Incidents Tracking System (WITS) Dataset
  - 3000 instances, 4 classes
- National White Collar Crime Center (NW3C) Dataset
  - 10,000 instances, 6 classes
- DNA Dataset
  - 3190 instances, 3 classes
- Nuclear Dataset
  - 4 isotopes, 302 instances each, 2 classes (present or absent)

Test results- formatting example

washington: Precision with 20.0% sample.

Feature set	NB	HONB	p-value
2-grams, 527	0.385325345417	0.37947574175	0.0406466390867
3-grams, 2000	0.40266389433	0.375343322958	0.014521465685
3-grams, 3000	0.373523926451	0.374325247295	0.903405524597
3-grams, 4000	0.3585433809	0.3743400866	0.0139980238447
4-grams, 2000	0.392484306005	0.380535335696	0.164834313783
4-grams, 3000	0.370731799773	0.37985090032	0.287833977719
4-grams, 4000	0.353818212397	0.381131790729	0.00259737910631
5-grams, 2000	0.379994917438	0.390275005497	0.171558979025
5-grams, 3000	0.358151589942	0.389511410926	0.000547733156007
5-grams, 4000	0.334608109035	0.391758488138	1.00583670218e-07
6-grams, 2000	0.370377583811	0.396809698708	0.00246463068078
6-grams, 3000	0.340214093766	0.395981397505	2.93861262261e-05
6-grams, 4000	0.316894209427	0.395805445225	1.24556271481e-07
7-grams, 2000	0.357872103688	0.389618807041	0.00031495210473
7-grams, 3000	0.331326868806	0.389539261212	3.50263997248e-06
7-grams, 4000	0.311107552954	0.389239632146	7.96397654677e-07

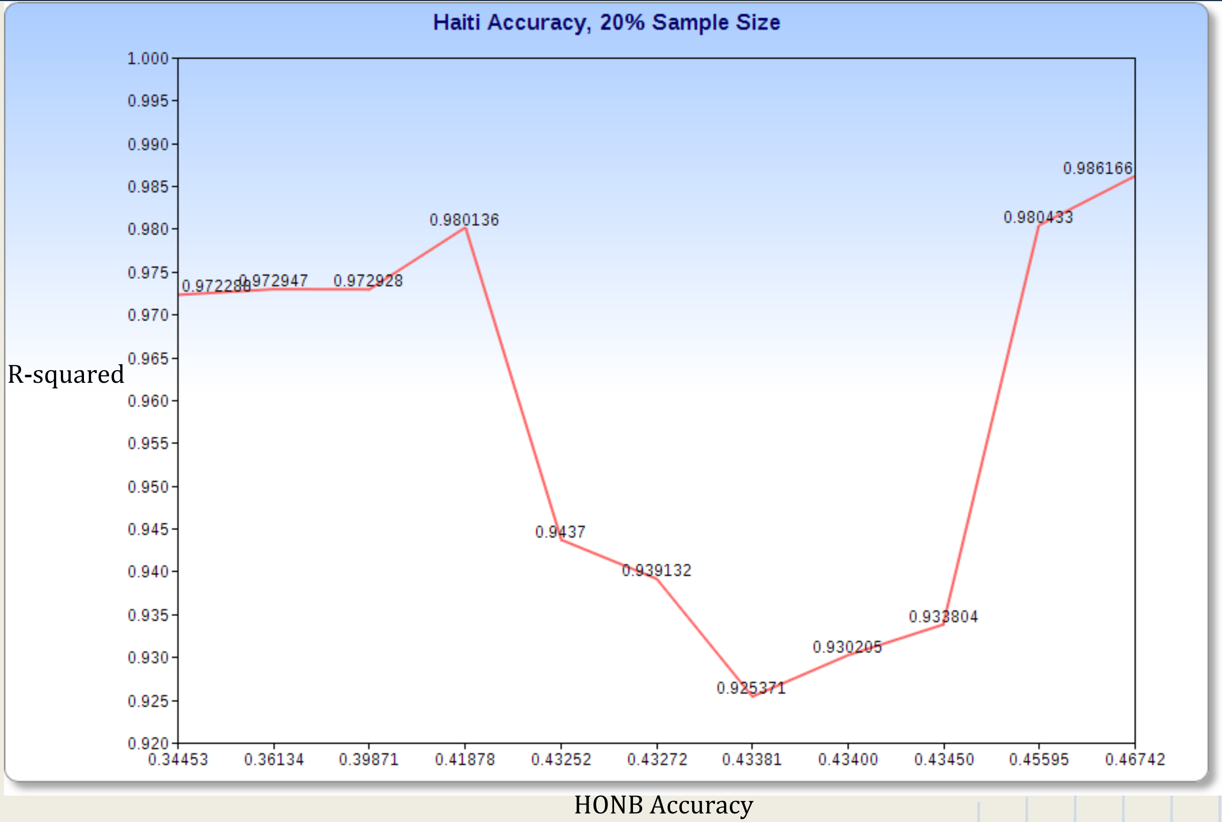
Background: Metrics

- Accuracy
- Precision
- Recall
- F-measure (F_β)
  - Harmonic mean of Precision & Recall
  - β -> relative weight

Results

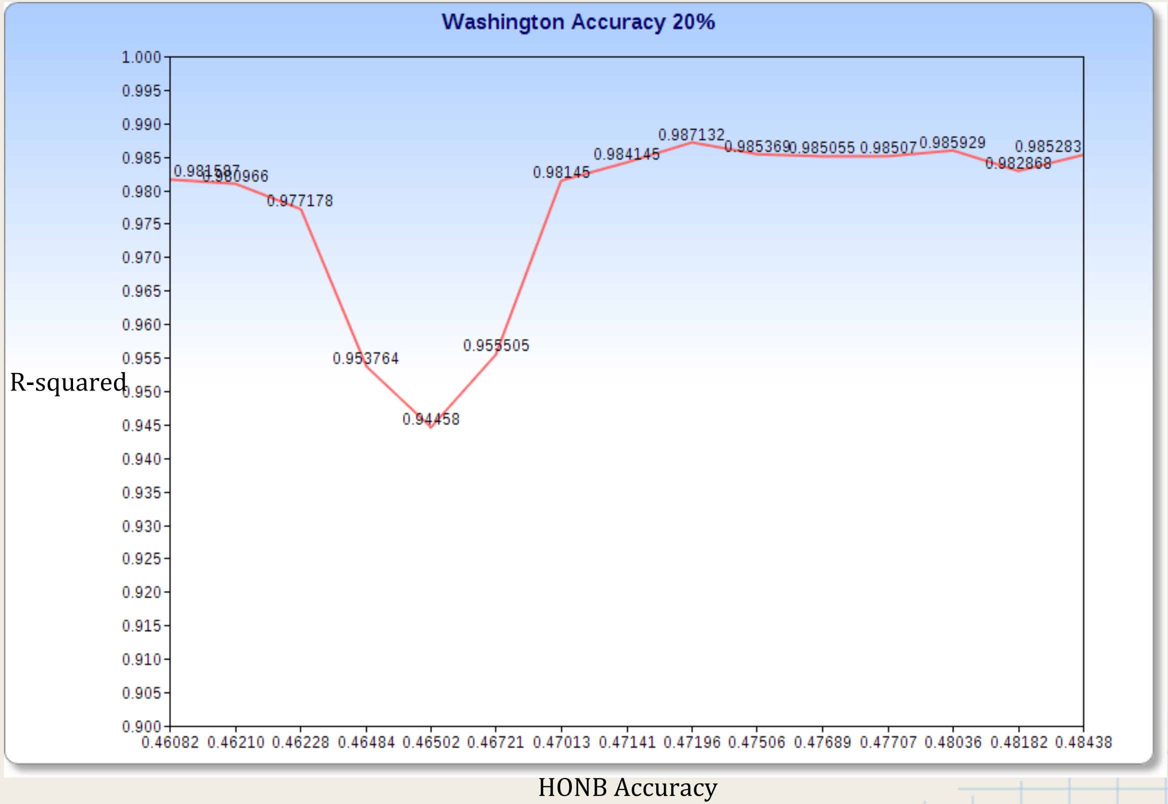
- Are our hypotheses supported by the evidence thus far?
  - A correlation seems to exist, but it’s not the type of correlation we were expecting
  - The nature of the correlation is not quite clear yet, but we are working on isolating it
- We therefore conclude that our original hypotheses are not supported by the experimental evidence, but a similar hypothesis may be

Results

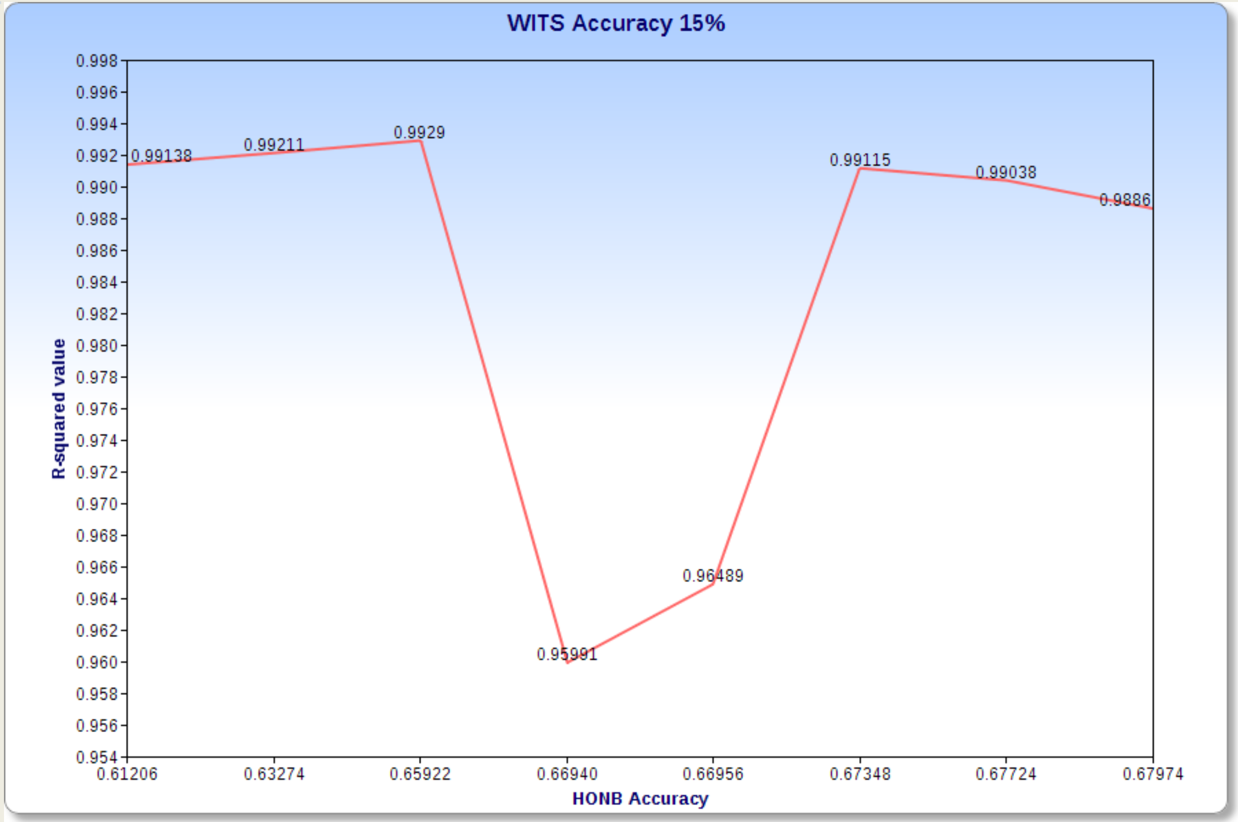


Haiti Accuracy, 20% Sample Size
R-squared
1.000
0.995
0.990
0.985
0.980
0.975
0.970
0.965
0.960
0.955
0.950
0.945
0.940
0.935
0.930
0.925
0.920
0.972288
0.972947
0.972928
0.980136
0.9437
0.939132
0.925371
0.930205
0.933804
0.980433
0.986166
0.34453
0.36134
0.39871
0.41878
0.43252
0.43272
0.43381
0.43400
0.43450
0.45595
0.46742
HONB Accuracy

Results


Washington Accuracy 20%
1.000
0.995
0.990
0.985
0.980
0.975
0.970
0.965
0.960
0.955
0.950
0.945
0.940
0.935
0.930
0.925
0.920
0.915
0.910
0.905
0.900
0.981597
0.980966
0.977178
0.953764
0.94458
0.955505
0.98145
0.984145
0.987132
0.985369
0.985055
0.98507
0.985929
0.982868
0.985283
0.46082 0.46210 0.46228 0.46484 0.46502 0.46721 0.47013 0.47141 0.47196 0.47506 0.47689 0.47707 0.48036 0.48182 0.48438
R-squared
HONB Accuracy

Results



WITS Accuracy 15%
R-squared value
HONB Accuracy
0.998
0.996
0.994
0.992
0.990
0.988
0.986
0.984
0.982
0.980
0.978
0.976
0.974
0.972
0.970
0.968
0.966
0.964
0.962
0.960
0.958
0.956
0.954
0.61206
0.63274
0.65922
0.66940
0.66956
0.67348
0.67724
0.67974
0.99138
0.99211
0.9929
0.95991
0.96489
0.99115
0.99038
0.9886

Future steps

- Investigate alternative ways of looking at results
  - Hold different variables constant: values of N, feature size, R-squared value, etc
- Create more data points
  - Compute results for more values of N and a wider range of feature spaces
- Pursue other possibilities for theorems
  - Might be able to say something about the performance of HONB in specific cases