

Increasing Effectiveness of HONB Microtext Classification in Disaster Response



BRIAN D. KNOPP
MENTORS: DR. WILLIAM M. POTTENGER AND
DR. CHRISTIE NELSON
DIMACS REU 2013

Background and Problem



- **Setting** – During disasters (earthquake, hurricane, etc.) rapid response is needed. On-the-ground situations are often communicated by *microtext*, short bits of text with some information content.
- **Problem** – Classifying text by hand is a slow and difficult process. Rapid response to people's needs is a high priority. A quick and accurate automated process is ideal.
- **Solution** – Leverage higher-order relationships between text attributes (words, groups of letters, etc).

Background – Ushahidi in Haiti and Chile

- Ushahidi is a volunteer disaster response group
 - Formed during the '07-'08 post-election violence in Kenya
 - Seen success in Haiti (2010), Chile (2010), and Middle East Democracy protests (2011)
- People send Ushahidi text

messages:

- Ex.: “THERES A BRIDGE IN DELMAS 31 THAT IS DAMANGED, THERES NO FOOD AND NO GAS YET”



Higher-Order Naïve Bayes



- Naïve Bayes is the traditional classifier due to its simplicity and efficiency
 - Relies on independent and identically distributed attributes
 - Real data (such as microtext) are not IID
- Information may be contained in higher-order relationships
- To utilize higher-order relationships HONB eliminates the IID assumption
 - Higher-order paths link attributes from different instances
- **Advantage** – HONB (like NB) classifies instances without needing additional contextual information
 - Explores latent connections between documents in a class

HONB - Paths



First Order Paths



Higher Order Paths



Number of Systems of Distinct Representatives



- Number of SDR is used in enumerating all higher-order paths
- Closed-form solutions are desirable for efficiency
 - Based on Inclusion-Exclusion principle of Set Theory
- Closed-form solutions for 2nd, 3rd, and 4th order:

○ For $S_I = |\bigcap_{i \in I} A_i|$ $\#SDR(A_1, A_2) = \sum_{i|j \in \lambda(1,1)} S_i S_j - \sum_{ij \in \lambda(2)} S_{ij}$

$$\#SDR(A_1, A_2, A_3) = \sum_{i|j|k \in \lambda(1,1,1)} S_i S_j S_k - \sum_{ij|k \in \lambda(2,1)} S_{ij} S_k + 2 \sum_{ijk \in \lambda(3)} S_{ijk}$$

$$\#SDR(A_1, A_2, A_3, A_4) = \sum_{i|j|k|l \in \lambda(1,1,1,1)} S_i S_j S_k S_l - \sum_{ij|k|l \in \lambda(2,1,1)} S_{ij} S_k S_l + 2 \sum_{ijk|l \in \lambda(3,1)} S_{ijk} S_l + \sum_{ij|kl \in \lambda(2,2)} S_{ij} S_{kl} - 6 \sum_{ijkl \in \lambda(4)} S_{ijkl}$$

My Investigation: Algorithm



- **Complexity**– The effect of #SDR on the algorithmic complexity of HONB
 - Obviously, increase #of sets will increase complexity.
 - What sort of benefit is gained by using a large # of sets?
 - Does a general closed form solution exist for all n sets?
 - ✦ Currently, only closed-form for $n \leq 4$
- **Accuracy** – May be able to increase classification accuracy by using different attributes
 - E.g.: Groups of letters other than words

My Project Goals and Methods



- Further investigate HONB
 - Background literature review
 - Review current program source-code
- Can classification be improved by using attributes other than whole words (or stemmed whole words)?
 - Create experimental methodology to determine
- What effect does the number of sets have on classification and complexity?
 - Is there any benefit to a large number of sets?
- Investigate computational complexity of the problem, given certain conditions
 - Number of documents to classify, number of classes, etc.
- If time, research closed-form for general n sets.

Citations



- Ganiz, M.C. *Higher-Order Path Analysis for Supervised Machine Learning*. Ph.D. dissertation, Lehigh University.
- Ganiz, M.C., George, C., Pottenger, W.M. *Higher Order Naïve Bayes: A Novel Non-IID Approach to Text Classification*. IEEE Transactions on Knowledge and Data Engineering.
- Nelson, C., Keiler, H., Pottenger, W.M. *Modeling Microtext with Higher Order Learning*.
- National Geographic. <http://news.nationalgeographic.com/news/2010/01/photogalleries/100114-aerial-haiti-earthquake-pictures>
- Ushahidi – Chile. chile.ushahidi.com

Thanks for your attention



ANY QUESTIONS?