Hierarchical Voting Experts: An Unsupervised Algorithm for Segmenting Hierarchically Structured Sequences

Matthew Miller
Department of Computer Science
Iowa State University
mamille@iastate.edu

Alexander Stoytchev
Department of Electrical and Computer Engineering
Iowa State University
alexs@iastate.edu

Introduction

This paper extends the Voting Experts (*VE*) algorithm (Cohen, Adams, & Heeringa 2007) to segment hierarchically structured sequences. The original algorithm was tested on text segmentation, and made use of two proposed characteristics of *chunks*, namely low internal entropy and high boundary entropy of segments. *VE* looks for these two properties, and uses them to segment sequences of tokens. It is surprisingly powerful given its simplicity, suggesting that the principle of segmenting based on low internal entropy and high boundary entropy is promising. Real world data often exhibits an inherently hierarchical structure, and it is well known that humans tend to chunk the world hierarchically (Miller 1956). It is therefore interesting to explore the applicability of a modified version of *VE* on hierarchically structured data.

We show that *VE* can be generalized to work on hierarchical data, and also that the higher order models can be used to improve the accuracy of the segmentation at lower levels.

Voting Experts

The *VE* algorithm (Cohen, Adams, & Heeringa 2007) consists of three main steps. Given a sequence for segmentation:

- Build and ngram trie of the sequence and use it to calculate the internal entropy and boundary entropy of each subsequence of length less than or equal to n, standardizing across all subsequences of the same length.
- Pass a sliding window of length n along the sequence. At each location, let each of two experts vote on how they would split the contents of the window - one minimizing the internal entropy of the two induced subsequences, the other maximizing the entropy at the split.
- Induce a split at each point in the sequence with a sufficient number of votes. Specifically look for "peaks" - locations with more votes than their neighbors.

For technical and implementation details of the algorithm, see the journal article (Cohen, Adams, & Heeringa 2007).

Hierarchical Voting Experts

The original application of *VE* was to segment text that had been stripped of punctuation and spaces. Accordingly, the ngram trie was built using the characters of the text. It is possible, however, to build a trie from any sequence of tokens upon which a total ordering can be imposed.

The extension to Hierarchical Voting Experts (*HVE*) is then natural. We run *VE* on a sequence to obtain a sequence of chunks. We then treat those chunks as the tokens of a sequence by imposing a lexicographical ordering on them. Then we run the algorithm again to obtain chunks of chunks. This process can be repeated indefinitely for any number of hierarchical layers (see Figure 1).

Additionally, we used the higher order model to increase the accuracy of the lower level segmentation. We ran standard *VE* on a given sequence to obtain a sequence of chunks. Then we built the second order model (ngram trie) using those chunks. Finally, we re-ran the algorithm on the original sequence with the addition of a third voting expert.

The third voting expert uses the higher order model to help split the lower order sequence. For each position of the sliding window, it checkes whether any subsequence of the window matches one or more chunks known to the higher order model. If so, it votes to place a break after the most common of those subsequences. If no match is found, the third expert does not vote. The reasoning is that, after building a higher order model, the most common tokens in that model will correspond to true common segments in the lower level sequence. So the third expert can recognize sequences that commonly become chunks, and vote to reinforce them.

Experimental Results

We designed several experiments to test the *HVE* algorithm. They demonstrate that the application of *HVE* to hierarchical data can induce accurate segmentation at each level. Additionally they show the effectiveness of the third voting expert technique on both hierarchical and non-hierarchical data.

Dataset: For all three experiments we used the first 50,000 characters of George Orwell's 1984 as our base text. This was one of the benchmark datasets used to evaluate the original *VE* algorithm, so it was chosen for comparison.

We used the base text to generate a hierarchically structured sequence. Specifically, we generated a mapping from

each alphanumerical character to a random three digit sequence. Then we used this mapping to translate the base text by replacing each character in it with its corresponding sequence. The digits in the random mapping ranged from 0 to 8. The three digit sequences were sufficiently similar to allow the translated dataset to be somewhat indeterminate. This introduced some error into the segmentation at the lower level.

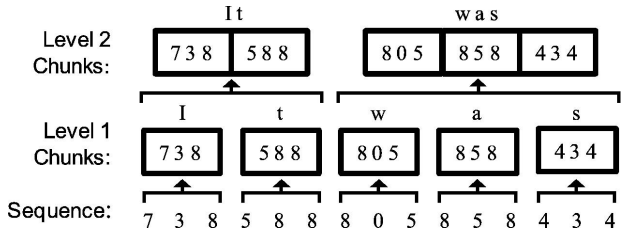

Figure 1: Illustration of hierarchical clustering

Experiment 1: We ran a two-level *HVE* on the hierarchical dataset, and evaluated the segmentation performance at each level. We used the f-measure, accuracy and hit rate of the induced boundary placements as our metric. These are the same metrics that were used to evaluate the *VE* algorithm. The results of the experiment are summarized in Table 1, along with the results obtained from running *VE* on the base text for comparison.

Experiment 2: We then applied the third voting expert technique at both levels of the hierarchy. The final results for both levels are also included in Table 1. We also ran the algorithm on subsets of the data, from 10% up to 100%, to illustrate its effectivness on smaller datasets. In figure 2 we compare the performance of the third voting expert technique with the standard *HVE* as the dataset increases in size. The graph focuses on the improvement of the second level segmentation.

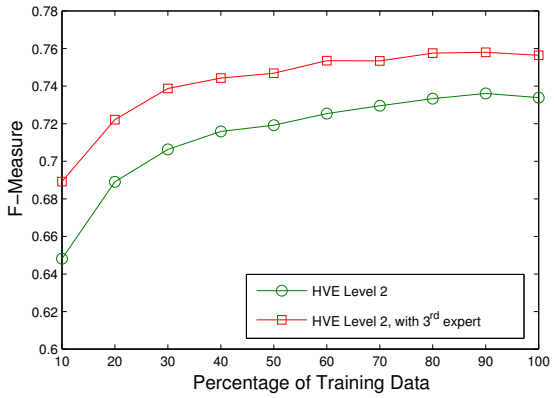

Figure 2: Performance as a function of the size of the data set. Notice the improvement due to the third voting expert.

Experiment 3: Finally we applied the third voting expert technique to the base text, which does not have a naturally hierarchical structure. However, it was still able to use the second order model to segment characters more accurately. The results are also summarized in Table 1.

Table 1: Performance Results for the *HVE* algorithm.

Test	F-measure	Accuracy	Hit Rate
Exp 1: Level 1	0.916	0.942	0.891
Exp 1: Level 2	0.734	0.748	0.720
Exp 1: base text	0.776	0.756	0.797
Exp 2: Level 1	0.945	0.965	0.926
Exp 2: Level 2	0.756	0.756	0.756
Exp 3: base text	0.784	0.752	0.818

Summary: It is clear that *HVE* is able to perform higher order segmentation of hierarchically structured data. In experiment 1, the word segmentation f-measure for the second level sequence is reasonably close to the baseline segmentation of words from the base text, despite the fact that the level 1 segmentation was not perfectly accurate, due to the added ambiguity of the three digit code used in the hierarchical data set. This shows that higher order segmentation is at least slightly robust to lower level segmentation error. However, it is not entirely robust, since the second level segmentation accuracy is seen to improve as the first level segmentation improves. Experiments 2 and 3 show us that we can indeed increase the accuracy of segmentation of both hierarchical and non-hierarchical sequences using the third voting expert technique.

Conclusions and Future Work

We have described a natural extension of the *VE* algorithm that can segment hierarchically structured data, and also the addition of a third voting expert to increase the segmentation accuracy using data from a higher order model. For further information and results, see http://www.public.iastate.edu/~mamille/AAAI08/.

The idea of segmentation based on low internal entropy and high boundary entropy has proven fruitful. We plan to explore several avenues for future improvement of the algorithm. Most immediately we want to make more efficient use of the higher order model to improve accuracy at the lower level. Additionally we would like to introduce the abstraction of sequences, to make the algorithm more robust to noise. And we would also like to build an online version that updates its model while it is segmenting. We would ultimately like to apply the algorithm to more difficult data like audio speech.

Acknowledgements

We would like to acknowledge Paul Cohen for giving us the source code for the original Voting Experts algorithm.

References

Cohen, P.; Adams, N.; and Heeringa, B. 2007. Voting experts: An unsupervised algorithm for segmenting sequences. *Journal of Intelligent Data Analysis*.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63:81–97.