Combining Feature Selectors for Text Classification

J. Scott Olsson
Dept. of Mathematics
University of Maryland
College Park, Maryland
olsson@math.umd.edu

Douglas W. Oard
College of Information Studies/UMIACS
University of Maryland
College Park, Maryland
oard@glue.umd.edu

ABSTRACT

We introduce several methods of combining feature selectors for text classification. Results from a large investigation of these combinations are summarized. Easily constructed combinations of feature selectors are shown to improve peak R-precision and F_1 at statistically significant levels.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: Miscellaneous

General Terms: Experimentation, Measurement

Keywords: text classification, feature selection

1. COMBINING FEATURE SELECTORS

We consider combinations of several widely used feature selection methods: document frequency thresholding, information gain, and the χ^2 methods $\chi^2_{\mathbf{max}}$ and $\chi^2_{\mathbf{avg}}$[8]. Each feature selector determines a rank ordering on the document's features (here words), sorted by score. These lists are taken as input to the combination methods. We consider combining two or more input methods using either the scores themselves or the rank ordering of the features.

In highest rank (**HR**) combination, we give as each feature's combined score the highest rank achieved in any of the input score lists. We expect this approach to be suitable if different input methods place different sets of informative features near the tops of their lists. This method discounts the negative information provided by an input; i.e., if one input has high confidence that a feature is uninformative, it will be overruled by any other input which ranks the feature higher. Highest rank combination has previously been used to combine hypothesized labels (rather than features) in classification problems [3]. Similarly, we consider combinations using the lowest (**LR**) and average (**AR**) rank.

We also consider combining by the feature selectors' normalized scores. Given two or more input vectors of feature scores, normalize them and, for each term, take the largest normalized feature score as the feature's combined score. We may normalize the input feature scores in several different ways. If we normalize input score vectors by dividing each element by the vector's largest element, then the OR effectively asks: for which input feature selection method did this term achieve a higher fraction of its largest observed score? We refer to this approach as **DMOR** (divide by maximum then OR). If on the other hand, we normalize input score vectors by dividing each element by the vector's (L_2) norm, then the OR asks: for which input method did this term achieve a higher fraction of the total achieved score across all the terms? We refer to this approach as **DLOR** (divide by length then OR). Previous feature selection combination studies [7] used a normalize then OR combination approach, although to our knowledge they were limited to pairwise combinations using DMOR.

In real world classification problems, low frequency terms are often dominated by noise. Many feature selection methods (particularly those of the χ^2 family) are known to be mislead by infrequent terms [1]. For these reasons, and because the Zipfian distribution of term frequencies leads to significant time savings through the elimination of low document frequency (DF) terms, text classification studies often disregard them. We call this practice DF *cutting*, and say that, if all terms with DF less than or equal to C are ignored, that we are cutting at level C. We consider combinations of feature selection methods computed at the same cutting level only.

2. EXPERIMENTS

We classify the documents using a local implementation k-Nearest Neighbors with symmetric Okapi term weighting [5]. For all trials, we fixed k at $k = 100$.

We conducted a preliminary study to explore the large space of possible combination methods. Using 23,149 documents from RCV1-v2 [4], we produce five partitions such that each training set contained 20% of the documents randomly assigned. On each of these splits, we ran every experimental setting (i.e., combination type, input methods, cutting level, and number of features). This entailed roughly 197 million classifications of documents.

Figure 1 shows R-precision averaged over the preliminary trials for each method investigated. The top few non-combination and combination methods are labeled. The best performing combination and non-combination methods are based on χ^2, with the best combinations using only two inputs. The non-combined methods appear to benefit from more aggressive cutting than the best combination methods. We see performance improves for the all-features case with low cutting, while it suffers if the cutting is too high.

We conducted a set of validation experiments, now including only those combination methods suggested by the previous study, but with many more trials to ensure statistical significance. Our validation data is a set of 200,000 new RCV1-v2 documents. We partitioned this new data into 20 disjoint sets of 10,000 documents each, before further divid-

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.
ACM 1-59593-433-2/06/0011.

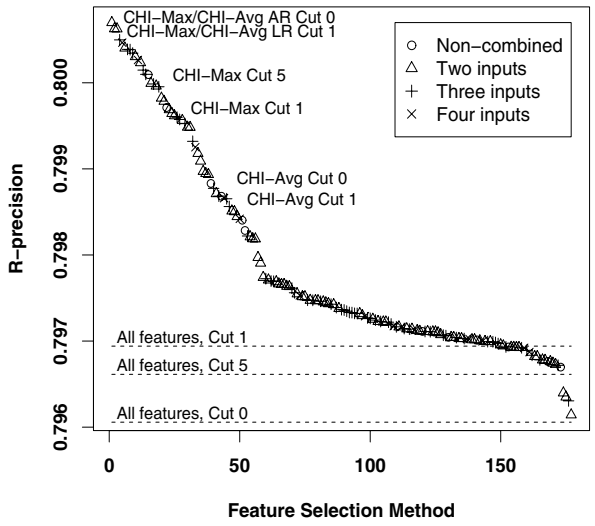

Figure 1: Peak R-precision averaged over the preliminary experiments for each method investigated. The top few non-combination and combination methods are labeled.

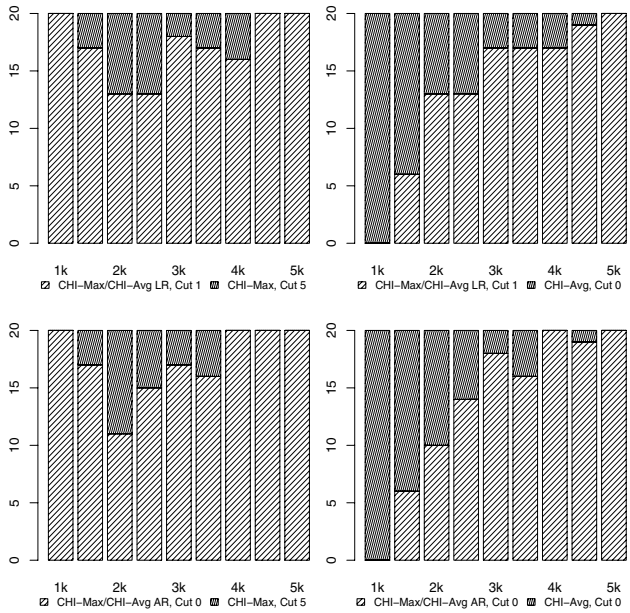

Figure 2: Bars depict how often each method achieved the higher R-precision in 20 trials, comparing combination and non-combination approaches.

Methods	R-precision	Micro-F_1
$\chi^2_{\mathbf{max}}/\chi^2_{\mathbf{avg}}$ AR vs. $\chi^2_{\mathbf{avg}}$	19 (2e-5)	18 (2e-4)
$\chi^2_{\mathbf{max}}/\chi^2_{\mathbf{avg}}$ AR vs. $\chi^2_{\mathbf{max}}$	16 (5.9e-3)	14 (5.8e-2)
$\chi^2_{\mathbf{max}}/\chi^2_{\mathbf{avg}}$ LR vs. $\chi^2_{\mathbf{avg}}$	18 (2e-4)	18 (2e-4)
$\chi^2_{\mathbf{max}}/\chi^2_{\mathbf{avg}}$ LR vs. $\chi^2_{\mathbf{max}}$	20 (9.5e-7)	17 (1.3e-3)

Table 1: How often combinations beat non-combinations for highest R-precision and F_1 in 20 disjoint test/train sets, considering all feature set sizes. Parentheses contain sign test p-values.

ing each of these in half to produce 20 pairs of testing and training sets of 5,000 documents. We again consider the 101 topic categories for our classification task.

We use the non-parametric Fisher Sign test for statistical significance in our improvements, comparing paired R-precisions for each of the 20 trials. For 20 trials, 15 wins is significant with p-value of 0.021. Figure 2 depicts how often each method in a pair achieved the highest R-precision, for each of the set sizes. We observe that the combination methods improve over non-combination methods at a statistically significant level over the majority of feature set sizes.

While combination methods achieved higher R-precision over the majority of individual set sizes, recall that we selected methods from our preliminary study to attain a highest *peak* R-precision. Table 1 shows how often each method in each pair obtained the highest peak over all set sizes. Because many complete tasks require thresholding to produce label sets, we also report here microaveraged F_1. We see that combination methods beat out the non-combination methods for peak R-precision, at statistically significant levels, for every pair observed. The improvement is similarly clear in the F_1 results, although we note that reversals occurred on a small number of the validation partitions.

3. CONCLUSIONS

We have seen that easily constructed combinations of feature selectors can achieve higher peak R-precision and microaveraged F_1 than their non-combined feature selection counterparts.

Future work might investigate the robustness of these combinations with respect to corpus characteristics (e.g., the skew in topic distribution) or the combination of other input feature selection methods (e.g., bi-normal separation [2]).

We refer the interested reader to [6] for a more detailed report of our investigation and results.

4. ACKNOWLEDGMENTS

This work was supported by NSF IIS 0122466 (MALACH).

5. REFERENCES

[1] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Comp. Ling.*, 19(1):61–74, 1993.

[2] G. Forman. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305, 2003.

[3] T. K. Ho, J. J. Hull, and S. N. Srihari. Decision Combination in Multiple Classifier Systems. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(1):66–75, 1994.

[4] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A New Benchmark Collection for Text Categorization Research. *J. Mach. Learn. Res.*, 5:361–397, 2004.

[5] J. S. Olsson. An analysis of the coupling between training set and neighborhood sizes for the kNN classifier. In *SIGIR '06*.

[6] J. S. Olsson and D. W. Oard. Exploring feature selection for multi-label text classification using ranked retrieval measures. University of Maryland CS technical report, *UMIACS-TR-2006-41*, 2006.

[7] M. Rogati and Y. Yang. High-performing feature selection for text classification. In *CIKM '02*.

[8] Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *ICML '97*.