High-Performing Feature Selection for Text Classification

Monica Rogati
CSD, Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
mrogati@cs.cmu.edu

Yiming Yang
CSD, Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
yiming@cs.cmu.edu

ABSTRACT

This paper reports a controlled study on a large number of filter feature selection methods for text classification. Over 100 variants of five major feature selection criteria were examined using four well-known classification algorithms: a Naive Bayesian (NB) approach, a Rocchio-style classifier, a k-nearest neighbor (kNN) method and a Support Vector Machine (SVM) system. Two benchmark collections were chosen as the testbeds: Reuters-21578 and small portion of Reuters Corpus Version 1 (RCV1), making the new results comparable to published results. We found that feature selection methods based on χ^2 statistics consistently outperformed those based on other criteria (including information gain) for all four classifiers and both data collections, and that a further increase in performance was obtained by combining uncorrelated and high-performing feature selection methods.

The results we obtained using only 3% of the available features are among the best reported, including results obtained with the full feature set.

Categories and Subject Descriptors

H.4 [**I.7**]: Document and Text Processing

General Terms

Experimentation, Verification, Performance

Keywords

text classification, feature selection

1. INTRODUCTION

Feature selection for text classification is a well-studied problem; its goals are improving classification effectiveness, computational efficiency, or both. Aggressive reduction of the feature space has been repeatedly shown to lead to little accuracy loss, and to a perfomance gain in many cases.

CIKM'02, November 4–9, 2002, McLean, Virginia, USA.

[12] conducted a comparative study on several feature selection criteria used in the filter model[4], and found χ^2 statistics and *information gain* more effective for optimizing classification results, and *document frequency* a better choice for efficiency and scalability if a small degradation in effectiveness is affordable. [5], [11] and [2] used hybrid approaches to address the dependency and redundancy among features. These algorithms require intensive calculations for the pairwise correlations between features, meaning a difficulty in scaling to a very large feature space.

A more efficient solution (tailored, however, to small 2-class collections), employed by [10], was to use information gain (assuming independence among features) to score and select an initial small, constant size subset of features, then enlarge the subset by computing the co-occurrences of the remaining features with regard to the selected ones. Other recent related work in feature selection and feature construction includes (but is not limited to) [1] (clustering words based on their cross-class distributions), [8],[9] etc.

Some important conclusions have not been reached yet, including

- Which feature selection methods are both computationally scalable and high-performing across classifiers and collections? Given the high variability of text collections, do such methods even exist?
- Would combining uncorrelated, but well-performing methods yield a performance increase?

This paper attempts to answer these questions by presenting a study of the performance of over 100 variants of 5 filter feature selection methods using two benchmark collections (Reuters-21578 and part of RCV1) and four classifiers (NB, Rocchio,KNN and SVM).

We also conducted experiments on combining high-performing but uncorrelated feature selection methods to improve classification results.

2. FEATURE SELECTION METHODS

We are concentrating on filter methods because 1) they are more scalable to very large collections and 2) their bias is different from the classifier's.

2.1 The Core Methods

We included several feature selection methods presented by [12]. They include document frequency (DF)(simply count the number of documents containing the feature), information gain (IG) (number of bits of information obtained for category prediction given a feature) and χ^2 (CHI) (measuring the lack of independence between a term and the category). [12] also used mutual information; due to its poor performance, we did not include this measure in our experiments. We did, however, include the binary version of information gain (IG2) because it is widely used.

2.2 Method Variations

There are several method variations we used:

- Using the term frequency instead of a binary value for each document counted in the scores (such variants would be identified by TF in the results; since none of these methods were among the top three performers, they do not appear on the graphs.)
- For the methods with one value per category (IG2, CHI,IG), we used both the average and the maximum value as the score. (identified by AVG, MAX)
- For IG and CHI, we also experimented with their generalized versions (combining evidence from all classes). They are identified by GEN.
- Eliminating rare words (DF ≤ 5) (identified by "cut")

2.3 Method Combinations

We examined the correlation between some of the top-performing methods and found that some (such as the multiclass version of IG and CHI_MAX) had little to negative correlation, which suggested a potential performance gain when they are combined. This was indeed the case for several methods (see "Results").

The combination of two methods was performed by normalizing the scores for each word and taking the maximum of the two scores (thereby performing an OR with equal weights given to the two methods being combined).

2.4 Redundancy Reducing Methods

We implemented a variant of the μ co-occurence method described by [10], which uses the other filter feature selection methods as a starting point.

While the complexity analysis in [10] is improved by the use of a tunable, arbitrary constant-size pool, we feel that using a percentage of the vocabulary is more adequate, since the size of the vocabulary can vary widely among collections. We implemented a variant of this method, using a percentage-based initial pool (1% instead of 5 terms) , smooth weighting instead of collection-dependent thresholding on the coocurrence and the multi-class version instead of 2-class.

As noted above, the total number of features, variants and combinations is well over 100. Each of the core methods has an average of approximately 3 variants, and the cca. 15 resulting methods were combined in pairs.

3. CLASSIFIERS AND DOCUMENT COLLECTIONS

We selected four high-performing classifiers for the feature selection experiments:

- K-Nearest Neighbors (local implementation)
- Naive Bayes (Rainbow, [7])
- Rocchio (local implementation)
- Support Vector Machines (SVMLight, [3])

In addition to Reuters-21578, we used a small percentage (1%) of the RCV1[6]. The "regional" class labels were selected for this task. In the remainder of the paper, this collection is identified as *RCV1-sampled*. The documents were chosen at random, and split into a 70% training set and 30% testing set. The resulting training set had 5000 documents and 198 categories. The next section shows results obtained with 5-fold cross-validation for all classifiers.

4. RESULTS

The following figures show the classifiers performance at several feature levels. While over 100 method variations have been tested (as described above), only the top 3 for each classifier are included, sorted by average performance over the [1-25] percent range.

A + sign indicates a combination of methods; a "cut" indicates that the low DF words were eliminated. Both macro F1 and micro F1 are presented for comparison purposes.

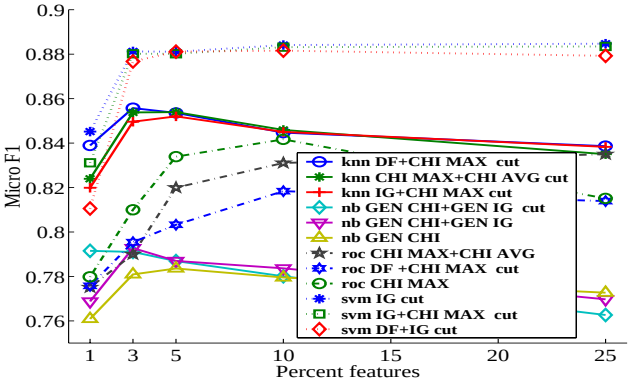

Figure 1: Top 3 feature selection methods for Reuters 21578 (Micro F1)

For Reuters-21578, our results are among the best reported in the literature for SVM, and consistent whith those in [12] for KNN.

SVM was also the least sensitive to different feature selection methods at the higher end of the performance spectrum included in the figure (top 3 methods). However, all classifiers had widely different performances across the feature selection methods we experimented with. Since, for clarity reasons, we only present the top three performing methods for each classifiers, the classifiers' behavior with regard to stability across the top three methods should not be generalized to all methods.

The main observations we would like to emphasize are:

- Almost all of the top performers had χ^2 as a component, regardless of classifiers. As pointed out in [12] χ^2 is normalized and scores are comparable across the same category.
- Eliminating the low DF words ("cut" runs) also boosted the performance; this step corrected the fact that χ^2 is known to be unreliable for rare words.
- Combining good methods with little or no correlation improved the results.

These observations also hold for macro-F1, as seen in Figure 4.

We observed that the methods which eliminate rare words (the "cut" methods) have better performance on Macro F1(rare categories) for *RCV1-sampled* than for Reuters 21578. One possible explanation for this behavior is that, in the "regional" taxonomy of *RCV1-sampled*, the rare words are more likely to be regionally informative terms (such as city names etc.) that are predictive of the regional class values, as opposed to noise and misspellings.

The "TF" set of methods did not appear in the top 5 methods in neither collection and for no classifier.

One difference worth noticing is that, while the results as measured by micro average F1 are highly clustered by classifiers, this does not hold for macro average F1. A good feature selection method enables KNN to surpass SVM's performance (see Figure 4).

We examined the effect of redundancy-reducing methods across classifiers, for aggressive feature selection (3%), where the best performance occured in previous experiments. The logarithm of the μ co-occurence was combined with the original feature selection method.

For Reuters-21578,the top method was unchanged(DF+CHI MAX). The μ-co-occurence only changed the ordering of

some methods at the lower performance levels of KNN.

The surprising performance boost came for Naive Bayes: the top 13 methods use the co-occurence information. The performance does not reach KNN's level. For *RCV1-sampled*, similar patterns were observed for KNN; NB and Rocchio, however, showed little improvement over using methods that do not utilize the cooccurence information.

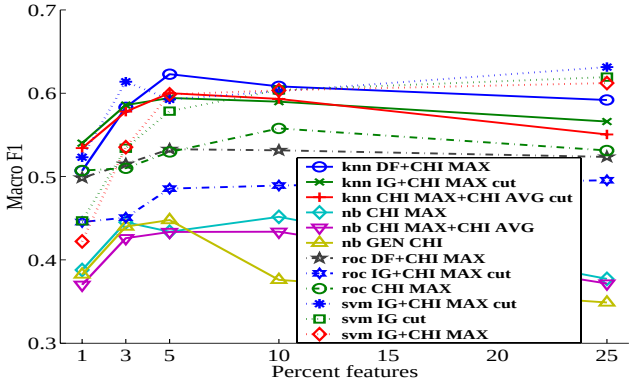

Figure 2: Top 3 feature selection methods for Reuters-21578 (Macro F1)

The observations above also hold for *RCV1-sampled*.

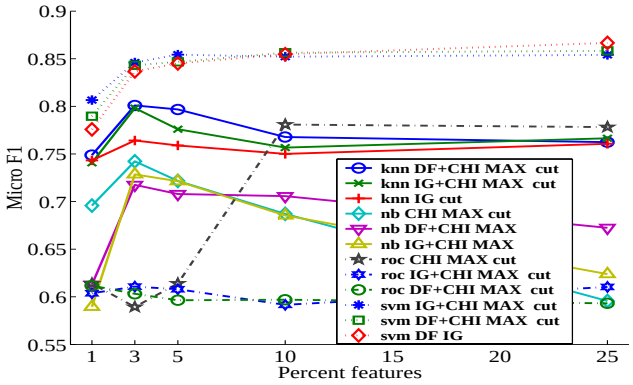

Figure 3: Top 3 feature selection methods for *RCV1-sampled* (Micro F1)

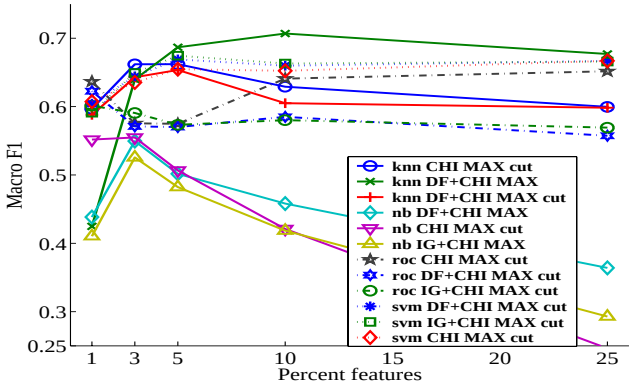

Figure 4: Top 3 feature selection methods for *RCV1-sampled*(Macro F1)

5. CONCLUSION

We conducted an extensive study of the performance of over 100 variants of 5 filter feature selection methods using two benchmark collections (Reuters 21578 and part of RCV1) and four classifiers (Naive Bayes, Rocchio, K-Nearest Neighbor and Support Vector Machines). The empirical results of our study suggest using filter methods which include the χ^2 statistic, combining them with DF or IG, and eliminating the rare words. Such methods were consistently better across classifiers, collections and performance measures.

We found that a redundancy-reducing method (using a modified version of μ-co-occurence) was computationally feasible; however, the results were not encouraging for aggressive feature selection at high performance levels, arguably because of the lack of proper weighting between the information-content scores and the redundancy scores.

6. ACKNOWLEDGMENTS

We thank Fan Li who helped correct an experimental error and generated some runs for the Rocchio classifier. This research is sponsored in part by National Science Foundation (NSF) under the grant number KDI-9873009, and in part by NSF under the grant number IIS-9982226. However, any opinions or conclusions in this paper are the authors' and do not necessarily reflect those of the sponsors.

7. REFERENCES

[1] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 96–103, 1998.

[2] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. In *International Conference on Machine Learning*, 2001.

[3] T. Joachims. Making large-scale support vector machine learning practical, 1998.

[4] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *International Conference on Machine Learning*, pages 121–129, 1994.

[5] D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

[6] T. Lewis, F. Li, R. Tony, and Y. Yang. The reuters corpus volume i as a text categorization test collection. 2002.

[7] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[8] J. Moore, E. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, and B. Mobasher. Web page categorization and feature selection using association rule and principal component clustering, 1997.

[9] A. Rozsypal and M. Kubat. Using the genetic algorithm to reduce the size of a nearest-neighbor classifier and to select relevant attributes. In *Proc. 18th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 2001.

[10] P. Soucy and P. Mineau. A simple feature selection method for text classification. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 897–902, 2001.

[11] E. P. Xing, M. I. Jordan, and R. M. Karp. Feature selection for high-dimensional genomic microarray data. In *Proc. 18th International Conf. on Machine Learning*, pages 601–608. Morgan Kaufmann, San Francisco, CA, 2001.

[12] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *International Conference on Machine Learning*, pages 412–420, 1997.