Combination of Feature Selection Methods for Text Categorisation

Robert Neumayer[1], Rudolf Mayer[2], and Kjetil Nørvåg[1]

[1] Norwegian University of Science and Technology,
Department of Computer and Information Science, Trondheim, Norway,
{neumayer,noervaag}@idi.ntnu.no
[2] Vienna University of Technology,
Institute of Software Technology and Interactive Systems, Vienna, Austria,
mayer@ifs.tuwien.ac.at

Abstract. Feature selection plays a vital role in text categorisation. A range of different methods have been developed, each having unique properties and selecting different features. We show some results of an extensive study of feature selection approaches using a wide range of combination methods. We performed experiments on 18 test collections and report a subset of the results.

1 Introduction

Feature selection is an essential technique to facilitate reduction in dimensionality which is a vital component of any text categorisation system. Indeed, most machine learning algorithms could not be applied at all without it. A range of methods have been suggested and evaluated to this end. A good overview and a comprehensive survey of the whole area is given in [4].

A recent and extensive empirical study of feature selection is performed in [1]. Here, the author compares a list of 11 (8 without modified methods) feature selection methods. The performance evaluation is done on 19 test collections of different size and difficulty. The author uses one-against-all classification and as such averages all results over 229 binary classification problems.

Feature selection combination was, for example, suggested in [3]. The authors selected feature selection methods based on 'uncorrelatedness' and presented results for two document collections. More experiments for text categorisation are reported in [2]. Experiments are done with four different feature selection methods and a test collection sampled from RCV1-v2. It is shown that certain combination methods improve peak R-precision and F_1. Both studies only partly work with benchmark collections and the results are difficult to compare.

2 Feature Selection Methods

We show the different feature selection methods we use in this paper in Table 1. If a method does not rely on previously assigned labels it is an unsupervised

method (methods belonging there are shown in the first part of the table), if it does it belongs to the category of supervised methods (shown in the second part of the table).

Table 1: Feature selection methods used throughout the paper

Method		Explanation
Document Freq.	(DF)	The number of documents a term occurs in.
Inverse Document Freq.	(IDF)	The inverse of the document Freq.
Collection Freq.	(CF)	The total number of occurrences of a term.
Inverse Collection Freq.	(ICF)	The inverse of the collection frequency.
Term Freq. Document Freq.	(TFDF)	A method based on thresholds for DF.
Information Gain	(IG)	A information theoretic method taking into account both negative and positive examples.
Mutual Information	(MI)	Another method from information theory.
Odds Ratio	(OR)	A probabilistic feature selection method.
Class Discrimination Value	(CDV)	OR variant targeted at multi-class problems.
Word Freq.	(WF)	The weighted number of occurrences per class.
χ^2 statistic	(χ^2)	Statistical method based on the independence of features.
NGL-Coefficient	(NGL)	A χ^2 variant only looking at positive examples.
Categorical Proportional Difference	(CPD)	Considers only positive examples.
GSS-Coefficient		Another simplified χ^2 method.
Bi-Normal Separation	(BNS)	Incorporates the inverse standard distribution and both positive and negative classes.

3 Combination Methods

In the following we show a range of ranking merging methods applicable to the problem of merging feature rankings generated by different methods. The available methods are listed in Table 2. The first part of the table lists method based on rank. The second and third part list methods based on value and on the round robin strategy, respectively.

Table 2: Ranking Merging methods used

Method		Explanation
Highest Rank	(HR)	A feature's highest rank in all single rankings.
Lowest Rank	(LR)	The lowest of all rankings is used as final score.
Average Rank	(AR)	The average over all single ranks is used.
Borda Ranking Merging	(BRM)	Gives scores according to the length of the single rankings.

Condorcet Ranking Merging	(CRM)	Is a majoritarian method favouring the candidate beating every other candidate in pair-wise comparisons.
Reciprocal Ranking Merging	(RRM)	In this setting, the final score for a feature is the sum of 1 divided by the rank in the single rankings.
Divide by Max. then OR	(DMOR)	The average over all single feature values in this setting we normalise by the maximum.
Divide by Length then OR	(DLOR)	Normalisation is performed via dividing by the length of the vector.
Pure Round Robin	(RR)	One feature is added from each ranking in turn until the desired number of features is reached.
Top N Ranking Merging	(Top$N$)	The top n features from each ranking in turn are added until enough features are collected.
Weighted N Ranking Merging	(WN)	The first n % are taken from the first ranking, the remaining $1 - n$ % are composed of the other rankings in equal parts.

4 Experiments

We used SVMs and five runs of four-fold cross validation. The results given are the macro averaged classification accuracies for both single methods and selected combinations. Based on the performance of the individual methods we chose the following combinations of feature selection methods (combined with the methods from Table 2 for our experiments: BNS, χ^2, DF, GSS, IG, MI, TFDF, WF and OR. This selection presents a good cross-section of the methods listed above since they both belong to different categories of methods and have show to have good performance in other studies in the past and show minimal to negative correlation with each other (based on both rank coefficient and classification performance).

We use a set of categorisation problems also used for binary classification experiments in [1], which were initially used by Han and Karypis. The collections were already preprocessed by basic stemming and stop-word removal. However, we use the sets for multi-class classification.

We show only a selection of all experiments. The accuracies for the top 200 selected features for both the single methods and combinations in Table 3. We chose to show results for 200 features because it is low enough so the classification is well possible. The best result per data set is shown in bold letters, the best result per method/combination in italic font. Overall we see that the combination methods outperform the single methods only in some cases, and never by much. On the other hand, the combinations are never much worse than the best single method. There is neither any single type of aggregation which provides the best results. However, for 100, 200, 500, and 1000 features, the method with the best averaged results is a combination method, even though the performance increase is very small.

Table 3. Average Classification Accurracies for the Top 200 Features

	BNS	χ^2	DF	GSS	IG	MI	OR	TFDF	WF	IG-BNS	IG-OR	OR-χ^2
la1	71.32	85.41	82.91	85.94	*86.27*	85.17	70.86	83.75	82.18	86.15	**86.39**	85.77
la12	71.60	87.29	85.38	87.55	*88.23*	86.73	69.87	85.48	83.85	88.23	**88.40**	86.94
oh0	86.20	*87.74*	67.10	85.24	86.44	86.38	82.41	75.91	68.57	86.44	86.66	**88.00**
oh10	**78.13**	77.70	67.18	76.52	77.96	77.11	74.53	73.26	69.85	*77.98*	*77.98*	77.77
oh15	79.15	**80.53**	62.98	78.25	79.39	78.29	70.58	72.75	63.64	79.41	79.59	*80.24*
oh5	**85.56**	84.47	74.68	85.03	84.10	84.34	81.68	79.35	79.26	84.12	84.18	*84.47*
ohsc	75.89	**77.87**	70.29	76.92	77.23	77.64	61.62	72.31	75.23	*77.22*	*77.22*	77.05
la2	70.67	86.64	83.64	87.57	*88.27*	86.13	73.40	84.34	82.43	88.26	**88.38**	87.19
wap	66.72	73.82	72.92	76.72	**80.83**	75.04	76.67	72.83	62.42	80.62	*80.81*	77.50
fbis	73.50	76.09	71.73	75.84	*82.83*	75.36	78.19	75.06	72.84	82.83	**82.90**	81.79
re1	83.68	85.23	74.29	84.49	*86.80*	83.32	78.94	78.20	73.82	86.76	**87.04**	86.13
tr11	85.07	**86.95**	83.77	86.13	86.43	85.75	84.54	85.31	83.00	86.33	*86.52*	86.42
tr12	85.11	82.74	73.87	80.64	*85.56*	83.89	79.17	77.70	70.10	85.49	**85.69**	84.83
tr21	80.24	87.92	83.57	89.40	*94.23*	83.81	86.96	84.52	82.44	93.81	94.23	**95.06**
tr23	64.51	80.39	83.53	82.25	*86.27*	81.86	86.27	82.94	74.71	86.18	86.78	**89.51**
tr31	95.84	95.83	92.68	95.58	**96.78**	95.36	93.69	95.64	91.54	*96.76*	96.57	95.60
tr41	92.71	*95.56*	88.68	94.65	95.03	94.94	91.82	91.78	87.79	95.03	95.15	**95.88**
tr45	86.00	91.42	83.77	90.84	*92.49*	91.36	87.39	86.03	81.34	92.38	92.93	**93.45**

5 Outlook and Future Work

We performed extensive feature selection and classification experiments on 18 different multi-class text categorisation problems. Further we used a wide range of ranking merging methods for combining features from multiple methods. However, no combination showed to be generally superior to the best single methods. Future work will deal with presenting more results in an accessible way and assessing the feasibility of ensemble methods to increase classification performance.

References

1. George Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.
2. J. Scott Olsson and Douglas W. Oard. Combining feature selectors for text classification. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06)*, pages 798–799, Arlington, Virginia, USA, November 6-11 2006. ACM.
3. Monica Rogati and Yiming Yang. High-performing feature selection for text classification. In *Proceedings of the 11th ACM International Conference on Information and Knowledge Management (CIKM'02)*, pages 659–661, McLean, Virginia, USA, November 4-9 2002. ACM.
4. Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.