Words as Rules: Feature Selection in Text Categorization[⋆]

E. Montañés, E.F. Combarro[⋆⋆], I. Díaz, J. Ranilla, and J.R. Quevedo

Artificial Intelligence Center, University of Oviedo, Spain
`ir@aic.uniovi.es`

Abstract. In Text Categorization problems usually there is a lot of noisy and irrelevant information present. In this paper we propose to apply some measures taken from the Machine Learning environment for Feature Selection. The classifier used is Support Vector Machines. The experiments over two different corpora show that some of the new measures perform better than the traditional Information Theory measures.

1 Introduction

Text Categorization (TC) [1] consists of assigning a set of documents to a set of categories. The removal of irrelevant or noisy features [2] improves the performance of the classifiers and reduces the computational cost.

The *bag of words* [1] is the most common document representation, using the absolute frequency (tf) to measure the relevance of the words over the documents [3]. *Stemming* and removing of *stop words* are usually performed. In this paper, words occurring in each category are used isolated from the rest [1] (local sets). The classification is tackled using the *one-against-the-rest* [4] approach and Support Vector Machines (SVM), since they perform fast and well [3] in TC.

This paper proposes some well-known impurity measures taken from the Machine Learning (ML) environment to perform Feature Selection (FS).

The rest of the paper introduces these measures, describes the corpora and the experiments and presents some conclusions and ideas for further research.

2 Feature Selection

FS is commonly performed in TC by keeping the words with highest score according to a measure of word relevance, like Information Theory (IT) measures. They consider the distribution of the words over the different categories. Some of the most adopted are *information gain* (IG) [5], *expected cross entropy for text* (CET) [6] and $S-\chi^2$ [7], a modification of the χ^2 statistic [2].

In the measures proposed here, fixed a category c, a word w is identified with the rule $w \rightarrow c$ which says: *If w is in a document, then the document belongs*

[⋆] This work has been supported under MCyT and Feder grant TIC2001-3579.

[⋆⋆] The author acknowledges the support of research project FICYT PR-01-GE-15.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3036, pp. 666–669, 2004.

to c. Then, the relevance of w for c is identified with the quality of the rule $w \to c$ [8].

Many popular rule quality measures are based on the percentage of successes and failures of the application of the rule. Two examples are the Laplace measure (L) and the *difference* (D) [9]. The former is a slight modification of the precision. The latter establishes a balance between the documents containing w and penalizes the words from documents not belonging to c. They are defined by

$$L(w \to c) = \frac{a_{w,c} + 1}{a_{w,c} + b_{w,c} + s} \qquad D(w \to c) = a_{w,c} - b_{w,c}$$

where $a_{w,c}$ is the number of documents of c in which w appears, and $b_{w,c}$ is the number of documents contaning w but not belonging to c.

We also propose variants that consider the absence of the word in c, penalizing more aggressively those words which appear in few documents of c. Hence, we define

$$L_{ir}(w \to c) = \frac{a_{w,c} + 1}{a_{w,c} + b_{w,c} + c_{w,c} + s} \qquad D_{ir}(w \to c) = a_{w,c} - b_{w,c} - c_{w,c}$$

where $c_{w,c}$ is the number of documents from c not containing w.

3 Experiments

Before presenting the experiments, we describe the corpora. Reuters-21578 contains short economic news. The distribution of the documents over the categories is quite unbalanced and the words are little scattered. Considering the Apté split [4] we obtain 7063 training documents and 2742 test documents, assigned to 90 different categories.

Ohsumed is a MEDLINE subset from 270 medical journals. Here, we consider the first $20,000$ MEDLINE documents from 1991 with abstract (the first $10,000$ for train and the rest for test) and the 23 subcategories of diseases (C of MeSH[1]). The words here are quite more scattered than in Reuters and the distribution of documents over the categories is much more balanced.

In the experiments, the filtering levels (fl) range from 20% to 98%. One-tail paired t-tests at significance level of 95% are conducted between $F_1's$ for each pair of measures. Table 1 presents the *macroaverage* and the *microaverage* of F_1 [1]. Tables 2 and 3 show the *t-test* results[2].

For Reuters, L_{ir}, D_{ir} and IG produce, in general, the best *macroaverage* and *microaverage* among their variants. This may be because the words are little scattered, that is, each category has a high percentage of specific words. Hence, the best measures are either those which tend to select words frequent in

---

[1] Medical Subject Headings `http://www.nlm.nih.gov/mesh/2002/index.html`

[2] In them, "+" means that the first measure is significantly better than the seconnd, "-" means that the seconnd measure is better than the first, "=" means that there exists no significative difference.

Table 1. *Macroaverage* and *Microaverage* of F_1 for different variants

Reuters														
	Macroaverage							*Microaverage*						
$fl(\%)$	L	L_{ir}	D	D_{ir}	CET	S-χ^2	IG	L	L_{ir}	D	D_{ir}	CET	S-χ^2	IG
20	47.15	47.53	46.41	46.71	46.92	46.35	46.07	82.71	85.03	80.99	81.54	83.39	83.35	84.78
40	48.31	48.90	45.17	45.75	46.90	47.27	48.03	81.53	85.16	79.42	80.49	83.03	83.19	84.96
60	48.40	49.01	44.06	44.61	46.89	46.67	47.84	80.16	85.26	78.91	79.79	83.03	83.28	85.02
80	43.73	48.66	42.41	43.77	48.55	48.87	48.57	77.69	85.21	78.68	79.99	83.33	83.40	85.42
85	43.74	48.40	41.61	44.06	48.84	47.69	48.44	76.80	85.05	78.63	80.25	83.40	83.17	85.53
90	41.64	48.09	40.30	42.48	48.12	47.87	48.69	75.26	84.81	77.77	80.26	82.99	83.30	85.48
95	33.45	47.42	38.87	41.42	47.61	47.82	49.19	74.15	84.47	75.63	79.26	83.22	83.57	85.40
98	30.43	45.73	37.85	40.68	47.55	46.30	48.41	74.76	83.43	78.40	80.37	83.02	83.01	84.76
Ohsumed														
	Macroaverage							*Microaverage*						
$fl(\%)$	L	L_{ir}	D	D_{ir}	CET	$S-\chi^2$	IG	L	L_{ir}	D	D_{ir}	CET	$S-\chi^2$	IG
20	46.27	42.49	49.54	50.00	45.92	45.86	42.71	53.53	51.51	56.19	56.77	53.46	53.47	51.93
40	51.11	42.91	50.57	51.52	46.73	46.71	43.61	56.92	51.54	56.20	57.25	53.91	53.94	52.31
60	51.32	43.66	50.01	51.11	47.39	47.28	45.21	56.63	51.99	55.52	56.89	54.49	54.35	53.13
80	48.42	44.31	48.77	50.74	48.14	47.40	46.94	53.41	51.90	55.11	57.18	54.80	54.49	53.98
85	46.40	44.14	48.93	51.17	48.36	47.64	47.11	51.32	51.24	53.45	57.54	55.10	54.75	54.19
90	47.46	44.57	50.26	52.19	48.79	47.85	47.54	52.22	51.15	54.36	57.73	55.42	54.93	54.59
95	48.33	44.17	51.58	52.37	48.91	46.94	49.40	53.75	49.94	55.72	58.04	55.69	54.48	55.47
98	47.82	41.02	52.27	51.44	47.47	44.16	49.66	53.08	45.85	56.24	57.00	54.04	51.99	54.81

Table 2. t-tests among different variants

	Reuters					Ohsumed				
$fl(\%)$	L_{ir}-L	D_{ir}-D	IG-CET	IG-S-χ^2	CET-S-χ^2	L_{ir}-L	D_{ir}-D	IG-CET	IG-S-χ^2	CET-S-χ^2
20	=	+	=	=	=	-	+	-	-	=
40	=	+	+	=	=	-	+	-	-	=
60	=	=	=	=	=	-	+	-	-	=
80	+	+	=	=	=	-	+	-	=	+
85	+	+	=	=	=	=	+	-	=	+
90	+	+	=	=	=	-	+	-	=	+
95	+	+	+	+	=	-	=	=	+	+
98	+	+	=	+	+	-	=	+	+	+

the category (like L_{ir} and D_{ir}) or those that consider the absence of the word in the category (like IG). Among them, L_{ir} and IG are statistically better, with no significative differences between them.

Regarding Ohsumed, L, D_{ir} and CET are, in general, the best measures among their variants, in *macroaverage* and in *microaverage*. Here, there are not so many specific words in each category, since the words are slightly scattered. Hence, the best measures are those that select words frequent in the category, although they appear in the rest, like L, D_{ir} and CET. Among them, D_{ir} is statistically better than the rest.

4 Conclusions and Future Work

This paper proposes some measures taken from Machine Learning for Feature Selection in TC, comparing them with other traditional Information Theory measures.

The performance of the measures depends on the corpus. For Reuters, which has an unbalanced distribution of documents and has the words little scattered, it

Table 3. *t-tests* among the best variants

Filtering level	20	40	60	80	85	90	95	98		20	40	60	80	85	90	95	98
Reuters									Ohsumed								
$L_{ir} - D_{ir}$	=	+	+	+	+	+	+	+	$L - D_{ir}$	-	=	=	-	-	-	-	-
$L_{ir} - IG$	+	=	=	=	=	=	-	=	$L - CET$	+	+	+	=	=	=	=	=
$D_{ir} - IG$	=	=	-	-	-	-	-	-	$D_{ir} - CET$	+	+	+	+	+	+	+	+

is better to penalize more those words which are not frequent in the category or to consider the absence of words in each category. For Ohsumed, whose distribution of documents is more uniform and which has the words highly scattered, it is better to reinforce the words of each category and not to penalize so much the words of the rest.

In our future work, we plan to use other classifiers and to find the optimal filtering levels for each measure, which may depend on properties of the category.

References

1. Sebastiani, F.: Machine learning in automated text categorisation. ACM Computing Survey **34** (2002)
2. Yang, T., Pedersen, J.P.: A comparative study on feature selection in text categorisation. In: Proceedings of ICML'97, 14th International Conference on Machine Learning. (1997) 412–420
3. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In Nédellec, C., Rouveirol, C., eds.: Proceedings of ECML-98, 10th European Conference on Machine Learning. Number 1398, Chemnitz, DE, Springer Verlag, Heidelberg, DE (1998) 137–142
4. Apte, C., Damerau, F., Weiss, S.: Automated learning of decision rules for text categorization. Information Systems **12** (1994) 233–251
5. Díaz, I., Ranilla, J., Montañés, E., Fernández, J., Combarro, E.: Improving performance of text categorization by combining filtering and support vector machines. (Journal of the American Society for Information Science and Technology (JASIST)) Accepted for publication.
6. Mladenic, D., Grobelnik, M.: Feature selection for unbalanced class distribution and naive bayes. In: Proceedings of 16th International Conference on Machine Learning ICML99, Bled, SL (1999) 258–267
7. Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the use of feature selection and negative evidence in automated text categorization. In: Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries, Lisbon, Portugal, Morgan Kaufmann (2000) 59–68
8. Combarro, E.F., Montañés, E., Ranilla, J., Fernández, J.: A comparison of the performance of svm and arni on text categorization whit new filtering measures on an unbalanced collection. In: International Work-conference on Artificial and Natural Neural Network, IWANN2003, Lecture Notes of Springer-Verlag. (2003)
9. Muggleton, S.: Inverse entailment and prolog. New Generation Computing, Special issue on Inductive Logic Programming **13** (1995) 245–286