Removing Redundancy from Relevant Features in Text Classification

E. Montañés
University of Oviedo
Spain
ir@aic.uniovi.es

I. Díaz
University of Oviedo
Spain
ir@aic.uniovi.es

E.F. Combarro
University of Oviedo
Spain
ir@aic.uniovi.es

J. Ranilla
University of Oviedo
Spain
ir@uniovi.es

Abstract

This paper proposes a method for Feature Selection in Text Categorization. This task is performed in two steps. Firstly, an analysis of relevance is performed and after that analysis of redundancy is done. For this purpose, a range of similarity measures are adopted and converted into symmetrical ones using several aggregation operators. This fact assures that the similarity between two words are independent of the order they are considered. Several experiments over four corpora are performed, leading to conclude that this method reaches good results.

Keywords: Redundancy, Feature Selection, Text Categorization, Information Retrieval.

1 Introduction

One of the main tasks in the processing of large collections of text files is that of assigning the documents of a corpus into a set of previously fixed categories, what is known as Text Categorization (TC) [16]. The most common way of representing the documents for TC is the bag of words (see [15]). This representation associates a vector to each document, where each component measures the importance of a certain word in it. This kind of domains involves a great amount of features most of them being irrelevant or redundant [15]. Under this circumstance, the classifier may be confused, may produce overfitting and increases its computational cost and the storage requirements. Thus, feature reduction often improves the effectiveness and efficiency of the classifier.

A common approach for feature reduction is Feature Selection (FS), which consists in choosing a subset of the original features for the document representation. Such selection could be performed evaluating each feature according to a scoring measure and keeping a predefined number of those with highest score [11, 17] or performing a subset evaluation search as hill-climbing or best first algorithm [10]. The first technique is more efficient but the features are purely obtained in terms of the relation between each feature and the target class. Hence, no dependence among features selected are taking into account, although empirical evidence shows that high dimensional domains also include redundant features. The second technique deals with both kinds of features, but redundant ones are implicitly handle with relevant ones and it is a not enough effective technique to deal with text domains. An efficient and effective alternative framework has been proposed in [18], which explicitly removes redundant features after relevant ones are selected. The work of this report is focused on this approach examining several ways of performing both tasks over some well know corpora for TC.

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 870–876
Torremolinos (Málaga), June 22–27, 2008

2 Previous Work

An alternative for FS is Feature Extraction (FE) methods, which transform or combine the original features to obtain a reduced number of features, like clustering [6] or Latent Semantic Indexing (LSI) [7]. These methods involve matrix management, which makes them be techniques with high computational cost. Hence, FS is generally more advisable for large domains as TC.

John et al. [9] distinguish two kinds of FS, namely filtering and wrapping. In the former, a feature subset is selected independently of the classifier. In the latter, a feature subset is selected using an evaluation function based on the classifier. Filtering approaches have been widely adopted, since wrapper ones usually result in a rather time consuming process.

Several measures have been proposed to quantify the relevance of features in TC. They range from simple ones taken from the Information Retrieval field, as document frequency [15] to those coming from Information Theory that consider the distribution of the words over the categories, such as, *information gain* [17]. Finally, rule quality measures [2, 12] are taken from the Machine Learning environment. These measures score a word w in a category c quantifying the quality of the rule that says *If a word w belongs a certain document, then such document belongs to category c.*

3 Measuring word/category and word/word closeness

The features in TC are the words appearing in the documents. This section describes the measures adopted in our approach for FS. In order to define them, let be c a category and w, w_1 and w_2 certain words. Then, for a pair (w, c) let define a as the number of documents of c where w appears and b as the number of documents containing w but not belonging to c. Also, for a pair (w_1, w_2) let define a as the number of documents containing both words and b as the number of documents containing w_1 but not w_2.

¿From those parameters, some measures coming from Information Retrieval (IR) field, as *document frequency* (*df*) are obtained. Other measures uses such parameters to compute the probabilities required for obtaining some Information Theory (IT) measures, like *information gain* (IG) [17] or *expected cross entropy for text* (CET) [11], which have obtained good results in TC. In case of Rule Quality (RQ) measures the rules denoted by $w \rightarrow c$ and $w_1 \rightarrow w_2$ are defined. The first one means that *If the word w appears in a document, then that document belongs to category c.* The second one means that *If the word* w_1 *appears in a document, then that document also contains* w_2. Then, the task of quantifying the relevance of w in c or the redundancy between w_1 and w_2 is converted into evaluating the quality of the corresponding rule [13]. The measures of this kind adopted here are *Laplace measure* (L), *difference* (D) and *impurity level* (IL). These measures and some variants of them (L_{ir}, D_{ir} and IL_{ir}) have been studied for FS in [13].

Linear and Angular measures arise from the study of the words that receive identical score under a measure by means of the level curves defined by the measure [3]. Certain selection of functions leads to Linear Measures (LM_k), mean other selection of functions yields Angular Measures (AM_k) [4], both depending on a real parameter k.

4 Symmetrical measures

The above measures are not symmetric. This property is desirable to quantify the closeness of two features, since it ensures that the closeness of w_1 to w_2 reaches the same score than the closeness of w_2 to w_1. Then, this paper proposes to take aggregation operators to convert such measures into symmetrical ones. The desirable aggregation operators Θ in this framework takes this definition

$$\begin{array}{l} \Theta: \quad \Re \quad x \quad \Re \quad \longrightarrow \quad \Re \\ (m_{1,2}, m_{2,1}) \rightarrow \Theta(m_{1,2}, m_{2,1}) \end{array} \qquad (1)$$

where $m_{1,2}$ and $m_{2,1}$ are the values of any of the measures that quantify the closeness of w_1 to w_2 and of w_2 to w_1 respectively.

There is a great variety of aggregation operators [5], although the common ones are the mean (Θ_{mean}), minimum (Θ_{min}) and maximum (Θ_{max}) which are the ones adopted here.

5 Performing relevant and redundant analysis

As commented in the introduction, this paper adopts the framework exposed in [18], where redundancy analysis is performed after relevance one. In this way, a ranking of the words according to the values $m(w, c)$ of a measure is obtained in first place. Then, some words from the bottom of the ranking could be removed, since they are the least relevant ones and hence could be considered irrelevant. Finally, redundant words are removed from the rest.

In [18], the redundant analysis is based on a greedy search which begins removing redundant features from the most relevant features to the least relevant ones. A feature is removed when it is possible to find an approximate Markov blanket for it formed by a predominant feature. A feature is a Markov blanket of other feature if it is more relevant and contains more information of the other feature than the category does ($m(w_1, w_2) > m(w_2, c)$). A feature is predominant if it does not have any approximate Markov blanket in the current set of features. A Markov blanket of a predominant feature is required since it avoids that a feature is removed after having been used for removing another previous one. It guarantees that a redundant feature removed earlier remains redundant when other features are removed later. In [18], they take symmetrical uncertainty both to compute $m(w_1, c)$ to estimate relevance and $m(w_1, w_2)$ to estimate redundancy. This process is shown in Figure 1

A new alternative consists in going from the least relevant word (the lowest value of $m(w, c)$) to the most relevant one. In this approach, the information of a word contained in the words more relevant than it is evaluated by means of a measure. This information is quantified individually for each more relevant word and then aggregated. The aggregation operators can be the same defined in the previous section, but must be defined over a set of real values instead of just over two real values. In this approach it is possible that a word previously used to remove a less relevant one could be also removed afterwards. But if this situation happens it means that it even exists a more relevant word which has previously made influence in its removal. The criterion adopted to consider a word as redundant is the same inequality as the Markov blanket approach. This new method differs from the above one in that it explores all features, mean the other one a feature removed in one stage is never considered for removing any of the rest. Figure 2 shows this process. More in detail, let be m a measure, Θ an aggregation operator, c a category, w_j a word and W_j the set of words more relevant than w_j according to m, that is

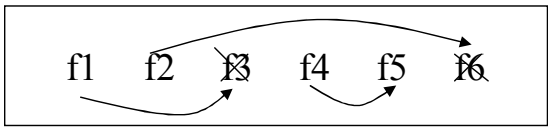

Figure 1: Approach based on Markov Blanket

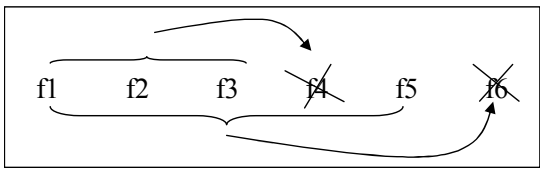

Figure 2: An alternative approach

$$W_j = \{w_k \quad s.t. \quad m(w_k, c) > m(w_j, c)\} \quad (2)$$

Then, the information W_j has of w_j is computed as $\Theta(m(w_1, w_j), ..., m(w_{|W_j|}, w_j))$ and compared to $m(w_j, c)$ to decide if w_j is discarded or not.

Table 1: *Microaverage* of F_1 when no feature reduction is previously applied

Reuters	Ohsumed	20news-bydate
$84,87\%$	$51,11\%$	$48,48\%$

Table 2: The best measure (m) from each group type (GT), *Microaverage* of F_1 and Filtering Level (FL) for Reuters with the best aggregation operator (mean)

MARKOV BLANKET			
No previous relevance filtering			
GT	m	F_1	FL
IRITRQ	IL_{ir}	$83,16\%$	83%
Linear	LM_7	$83,00\%$	92%
Angular	AM_{10}	$80,04\%$	95%
50% **previous relevance filtering**			
GT	m	F_1	FL
IRITRQ	IL_{ir}	$83,30\%$	85%
Linear	LM_7	$82,95\%$	92%
Angular	AM_{10}	$80,30\%$	97%
SYMMETRICAL MEASURES			
No previous relevance filtering			
GT	m	F_1	FL
IRITRQ	D	$84,06\%$	61%
Linear	LM_1	$84,06\%$	61%
Angular	AM_1	$83,78\%$	87%
50% **previous relevance filtering**			
GT	m	F_1	FL
IRITRQ	L_{ir}	$83,46\%$	78%
Linear	LM_9	$83,11\%$	77%
Angular	AM_1	$83,99\%$	89%
ALTERNATIVE APPROACH			
No previous relevance filtering			
GT	m	F_1	FL
IRITRQ	L	$85,35\%$	46%
Linear	LM_2	$85,26\%$	44%
Angular	AM_1	$85,20\%$	75%
50% **previous relevance filtering**			
GT	m	F_1	FL
IRITRQ	IL_{ir}	$84,51\%$	66%
Linear	LM_7	$83,14\%$	59%
Angular	AM_1	$85,07\%$	84%

6 Experiments

Three corpus have been taken for the experiments. One is the Apté split [1] in test and train documents of economic news Reuters-21578 published by Reuters in 1987. Ohsumed corpus is a clinically-oriented MEDLINE, whose the first 10000 documents of 1991 have been labelled as training documents and the following 10000 as testing documents. They have been split in [8] into the 23 subcategories of diseases of MeSH [14]. Finally, the collection 20news-bydate is formed by 20 categories of documents taken from the Usenet newsgroups collection.

The measures exposed in section 3 have been used both in relevance and redundancy stages and they were joined in three groups for the experiments. Let denote by the IRITRQ group the set of IR measures, IT measures and RQ measures. Similarly, the linear measures with the parameter k ranging from 1 to 10 form the Linear group and the Angular group contains the angular measures with the parameter k ranging from 1 to 10. The Θ_{mean}, Θ_{min} and Θ_{max} have been taken as aggregation operators for converting the measures into symmetrical ones and for the aggregation required in the alternative approach. Two set of experiments have been carried out. The first one uses relevance analysis just to produce a ranking, mean in the second one the 50% of the words in the ranking have been removed before performing redundant analysis. For each set, Markov blanket approach with original measures, Markov blanked approach converting original measures into symmetrical ones and the alternative approach have been compared. The performance was evaluated according to microaveraged F_1 [16].

Table 1 presents the performance of the classification when no feature reduction is previously applied, which will be considered as reference.

Table 2 shows the microaverage F_1 and the filtering level (FL) obtained for the best measure of each group when applying the techniques over Reuters. It only shows the re-

Table 3: The best measure (m) from each group type (GT), *Microaverage* of F_1 and Filtering Level (FL) for Ohsumed with the best aggregation operator (maximum)

MARKOV BLANKET			
No previous relevance filtering			
GT	m	F_1	FL
IRITRQ	IL	$57,71\%$	76%
Linear	LM_2	$57,57\%$	86%
Angular	AM_9	$54,79\%$	86%
50% **previous relevance filtering**			
GT	m	F_1	FL
IRITRQ	IL	$57,65\%$	76%
Linear	LM_2	$57,40\%$	86%
Angular	AM_9	$54,69\%$	89%
SYMMETRICAL MEASURES			
No previous relevance filtering			
GT	m	F_1	FL
IRITRQ	IL	$57,38\%$	77%
Linear	LM_2	$57,29\%$	86%
Angular	AM_9	$54,70\%$	87%
50% **previous relevance filtering**			
GT	m	F_1	FL
IRITRQ	IL	$57,27\%$	77%
Linear	LM_2	$57,20\%$	86%
Angular	AM_9	$54,74\%$	89%
ALTERNATIVE APPROACH			
No previous relevance filtering			
GT	m	F_1	FL
IRITRQ	L	$57,18\%$	76%
Linear	LM_1	$56,91\%$	86%
Angular	AM_1	$58,50\%$	93%
50% **previous relevance filtering**			
GT	m	F_1	FL
IRITRQ	IL_{ir}	$57,29\%$	94%
Linear	LM_7	$57,29\%$	94%
Angular	AM_1	$58,50\%$	93%

Table 4: The best measure (m) from each group type (GT), *Microaverage* of F_1 and Filtering Level (FL) for 20news-bydate with the best aggregation operator (mean)

MARKOV BLANKET			
No previous relevance filtering			
GT	m	F_1	FL
IRITRQ	CET	$61,62\%$	67%
Linear	LM_5	$61,56\%$	84%
Angular	AM_8	$55,97\%$	90%
50% **previous relevance filtering**			
GT	m	F_1	FL
IRITRQ	CET	$61,59\%$	71%
Linear	LM_5	$61,55\%$	84%
Angular	AM_8	$55,86\%$	91%
SYMMETRICAL MEASURES			
No previous relevance filtering			
GT	m	F_1	FL
IRITRQ	L	$62,19\%$	54%
Linear	LM_6	$61,45\%$	64%
Angular	AM_1	$60,53\%$	76%
50% **previous relevance filtering**			
GT	m	F_1	FL
IRITRQ	D_{ir}	$61,99\%$	75%
Linear	LM_2	$61,99\%$	75%
Angular	AM_1	$60,63\%$	80%
ALTERNATIVE APPROACH			
No previous relevance filtering			
GT	m	F_1	FL
IRITRQ	L_{ir}	$58,67\%$	20%
Linear	LM_{10}	$49,13\%$	1%
Angular	AM_{10}	$58,12\%$	74%
50% **previous relevance filtering**			
GT	m	F_1	FL
IRITRQ	D_{ir}	$62,01\%$	50%
Linear	LM_5	$62,27\%$	50%
Angular	AM_{10}	$57,14\%$	93%

sults for the best operator which is this case was the Θ_{mean} one. Only the Alternative Approach beats the reference, especially when no previous relevance feature selection is carried out. The best F_1 is reached by the alternative approach using as aggregation operator and the L measure. In general the Alternative Approach offers better results than the Symmetrical Markov Blanket approach, and in turn, this last one outperforms the Markov Blanket one. IRITRQ measures offer better results in any case and no previous relevance filtering seems to be the best option. Notice that the FL obtained by Markov Blanket are much more aggressive than those offered by the Alternative Approach.

The microaverage F_1 and the FL obtained for the best measure of each group when applying the techniques over Ohsumed is presented in Table 3, but in this case the Θ_{max} aggregation operator is taken because it leads to the best F_1. In this case, all approaches significantly beat the reference, but again the Alternative Approach reaches the best performance. In this case it does it using angular measure AM_1 with an aggressive FL. In this corpus, it hardly exists differences between performing a previous relevance filtering and performing none. It also seems that Markov Blanket produce slightly better F_1 than Symmetrical Markov Blanket.

Finally, Table 4 presents the microaverage F_1 and the FL produced by the best measure of each group when applying the techniques over 20news-bydate. For this corpus the Θ_{mean} is the aggregation operator which offers best results. As for Ohsumed, all approaches far improve the reference. The linear measure LM_5 in the Alternative Approach reaches the best F_1, although similar F_1 is obtained using L in the Symmetrical Markov Blanked approach. Only the Alternative Approach seems to improve F_1 when a previous relevance filtering is performed. In any case, the best results correspond to moderate FL.

7 Conclusion

This work focuses on performing analysis relevance before analysis redundancy for Feature Selection in Text Categorization. It studies the behaviour of a large range of scoring measures for this purpose, converting them into symmetrical ones by introducing aggregation operators and comparing the Markov blanket approach with a new alternative based on scanning the features from the least relevant ones to the most relevant ones and using aggregation operators.

Several experiments have been carried out over some corpora concluding than the alternative approach with certain measures and certain aggregation operators slightly improves the rest of approaches and that in general a previous relevance filtering does not lead to an improvement, otherwise in several situations the results are worst.

Acknowledgements

This research has been partially supported by a MEC and a FEDER grant TIN2007-61273.

References

[1] C. Apte and F. Damerau and S. Weiss (1994). Automated Learning of Decision Rules for Text Categorization. In *Information Systems*, volume 12, number 3, pages 233-251.

[2] E. F. Combarro and E. Montañés and J. Ranilla and J. Fernández (2003). A Comparison of the Performance of SVM and ARNI on Text Categorization whit New Filtering Measures on an Unbalanced Collection *Proc. International Work-Conference on Artificial and Natural Neural Network IWANN2003*, volume 2687, pages 743-749, Menorca, Spain.

[3] E.F. Combarro and E. Montañés and I. Díaz and J. Ranilla and R. Mones (2005). Introducing a Family of Linear Measures for Feature Selection in Text Categorization. In *IEEE Transactions on Knowl-*

edge and Data Engineering, volume 17, number 9, pages 1223-1232.

[4] E.F. Combarro and E. Montaés and J. Ranilla and I. Díaz (2006). Angular Measures for Feature Selection in Text Categorization. In *21st Annual ACM Symposium on Applied Computing SAC2006*, volume 1, pages 826-830, Dijon, France, Abril 2006.

[5] M. Detyniecki. Mathematical Aggregation Operators and Their Application to Video Quering. In *M.Sc. Thesis, Computer Science Division*. University of California, Berkeley, USA.

[6] Inderjit S. Dhillon and Subramanyam Mallela and Rahul Kumar (2003). A divisive information theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, volume 3, pages 1265-1287.

[7] G. Forman (2003). An extensive empirical study of Feature Selection Metrics for Text Categorization. In *Journal of Machine Learning Research*, volume 3, pages 1289-1305.

[8] T. Joachims (1998). Text categorization with support vector machines: learning with many relevant features. In *Proc. 10th European Conference on Machine Learning ECML98*, number 1398, Springer-Verlag, pages 137-142, Chemnitz, DE.

[9] G.H. John and R. Kohavi and K. Pfleger (1994). Irrelevant Features and the Subset Selection Problem. In *Proc. 11th International Conference on Machine Learning ICML94*, pages 121-129.

[10] R. Kohavi and G.H. John (1997). Wrappers for Feature Subset Selection. In *Artificial Intelligence*, volume 97, number 12, pages 273-324.

[11] D. Mladenic and M. Grobelnik (1999). Feature Selection for Unbalanced Class Distribution and Naive Bayes. In *Proc. 16th International Conference on Machine Learning ICML99*, pages 258-267, Bled, SL.

[12] E. Montañés and J. Fernández and I. Díaz and E. F. Combarro and J. Ranilla (2003). Measures of Rule Quality for Feature Selection in Text Categorization. *Proc. 5th International Symposium on Intelligent Data Analysis Berlin IDA2003*, volume 2810, pages 589-598, Berlin, Germany

[13] E. Montañés and I. Díaz and J. Ranilla and E.F. Combarro and J. Fernández (2005). Scoring and Selecting Terms for Text Categorization. In *IEEE Intelligent Systems*, volume 20, number 3, pages 40-47.

[14] National Library of Medicine (1993). Medical Subject Headings (MeSH). In *www.nlm.nih.gov/ mesh/ 2002/ index.html*, Bethesda, USA.

[15] G. Salton and M. J. McGill (1983). An introduction to modern information retrieval. McGraw-Hill.

[16] F. Sebastiani (2002) Machine Learning in Automated Text Categorisation. *ACM Computing Survey*, volume 34, number 1.

[17] Y. Yang and J. O. Pedersen (1997). A Comparative Study on Feature Selection in Text Categorisation. In *Proc. 14th International Conference on Machine Learning ICML97*, pages 412-420.

[18] L. Yu and H. Liu (2004) Efficient Feature Selection via Analysis of Relevance and Redundancy. In *Journal of Machine Learning Research*, volume 5, pages 1205-1224.