Finding Optimal Linear Measures for Feature Selection in Text Categorization

Elena Montañés  
Artificial Intelligence Center  
University of Oviedo  
Campus de Viesques S/N  
Gijón, Spain  
elena@aic.uniovi.es

E.F. Combarro  
Artificial Intelligence Center  
University of Oviedo  
Campus de Viesques S/N  
Gijón, Spain  
elias@aic.uniovi.es

José Ranilla  
Artificial Intelligence Center  
University of Oviedo  
Campus de Viesques S/N  
Gijón, Spain  
ranilla@uniovi.es

ABSTRACT
A common way of performing Feature Selection in Text Categorization consists in keeping the features with highest score according to certain measures, like linear ones which have been successfully proposed in [1]. Its disadvantage is that they need to previously determine the parameter which defines them. Until now, this drawback has been overcome by taking manually a set of values for such parameter. This paper proposes a method for automatically determining optimal values of the parameter by means of solving a univariate maximization problem.

Categories and Subject Descriptors
I.5.2 [Pattern Recognition]: Design Methodology—Feature evaluation and selection; I.7.1 [Document and Text Processing]: Document and Text Editing—Document management

General Terms
Theory, Measurement, Experimentation, Performance

1. INTRODUCTION
In Text Categorization (TC) [6] the documents are commonly represented using the bag of words and the absolute frequency (tf) of the word [2, 8]. It is also common to carry out the classification adopting the one-against-the-rest approach and to use Support Vector Machines (SVM) with linear kernel [2] as classifier [7]. The effectiveness of the classification use to be quantified by the macroaverage and the microaverage [6].

The main problem is that the documents are normally represented by a great number of features and most of them could be irrelevant [2]. As first approach, it is usually to remove stop words and to perform stemming. But, an additional Feature Selection (FS) usually helps to improve the performance of the classifiers and to reduce the computational time and the storage requirements.

A common way of tackling FS in TC consists in scoring the features using a certain measure, ordering them according to this measure and keeping or removing a predefined number or percentage of them. Several measures have been proposed for this purpose, like information gain (IG) [8], expected cross entropy for text (CET) [3] or like modified Laplace ($L_{mp}$) [4] or modified difference ($D_{mp}$) [4]. Recently, a new family of measures, called Linear Measures (LM), have also been proposed [1], but until now the parameter with which they are defined has to be manually selected. This paper proposes a method to automatically determine optimal values of such parameter.

2. OPTIMAL LINEAR MEASURES
Before defining the LM measures, let identify a word $w$ of a category $c$ by the pair $(a_{w,c}, b_{w,c})$, being $a_{w,c}$ the number of documents belonging to category $c$ in which $w$ appears and $b_{w,c}$ the number of documents in which $w$ occurs but not belonging to category $c$. Hence, in [1] the family of LM were exhaustively deducted from their level curves leading to the following expression

$$LM_k(w) = ka_{w,c} - b_{w,c}$$

with $k$ any real number which defines them.

A proposal is to convert the problem of finding an optimal value $k^*$ for the family $LM_k(w)$ $k \geq 0$ into an univariate maximization problem with $F_1$ as target function. Formally,

$$\text{Find } k^* \in [0, +\infty) \text{ such that } F_1(k^*) \geq F_1(k), \forall k \in [0, +\infty)$$

As it was stated in [1], only values of $k$ such that $k_{\min} \leq k \leq k_{\max}$ are interesting, where $k_{\min} = \min_{(a_{w,c}, b_{w,c})} (b_{w,c} / a_{w,c})$ and $k_{\max} = \max_{(a_{w,c}, b_{w,c})} (b_{w,c} / a_{w,c})$ for all the words $w$ of category $c$ with $(a_{w,c}, b_{w,c})$.

There exist several methods to tackle univariate maximization [5], but unfortunately, not all of them are able to be applied to our problem, since our target function ($F_1$) is not regular enough. The assumption of being continuous does...
of old defined as the minimum difference between the values proposed consists in automatically obtain a thresh-

The evaluation function was the target of automatically defining an optimal value of \( k \). Validation results reported in \cite{1} (LM Search, Uniform Search and Dicotomic Search methods. Not even hold, since it is a step function. Hence, only some

Table 1: Macroaverage and Microaverage of \( F_1 \) for Reuters-21578

<table>
<thead>
<tr>
<th>( IG )</th>
<th>( L_{18} )</th>
<th>( LM_{60} )</th>
<th>( LM_{90} )</th>
<th>( \text{Fib.} )</th>
<th>( \text{Unif.} )</th>
<th>( \text{Dic.} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>42.96</td>
<td>42.96</td>
<td>42.96</td>
<td>42.96</td>
<td>42.96</td>
<td>42.96</td>
</tr>
<tr>
<td>20</td>
<td>46.07</td>
<td>47.53</td>
<td>47.42</td>
<td>46.59</td>
<td>46.87</td>
<td>48.32</td>
</tr>
<tr>
<td>40</td>
<td>48.03</td>
<td>48.90</td>
<td>49.00</td>
<td>46.56</td>
<td>46.87</td>
<td>48.56</td>
</tr>
<tr>
<td>60</td>
<td>47.84</td>
<td>49.01</td>
<td>49.41</td>
<td>46.14</td>
<td>47.19</td>
<td>48.84</td>
</tr>
<tr>
<td>80</td>
<td>48.57</td>
<td>48.66</td>
<td>49.24</td>
<td>47.85</td>
<td>49.14</td>
<td>48.79</td>
</tr>
<tr>
<td>100</td>
<td>48.44</td>
<td>48.40</td>
<td>50.70</td>
<td>47.71</td>
<td>48.25</td>
<td>49.96</td>
</tr>
<tr>
<td>120</td>
<td>48.69</td>
<td>48.09</td>
<td>49.08</td>
<td>48.01</td>
<td>48.67</td>
<td>49.43</td>
</tr>
<tr>
<td>140</td>
<td>49.19</td>
<td>47.42</td>
<td>48.35</td>
<td>47.12</td>
<td>48.24</td>
<td>49.20</td>
</tr>
<tr>
<td>160</td>
<td>48.41</td>
<td>45.73</td>
<td>47.42</td>
<td>46.03</td>
<td>47.41</td>
<td>48.31</td>
</tr>
</tbody>
</table>

Table 2: Macroaverage and Microaverage of \( F_1 \) for Ohsumed

<table>
<thead>
<tr>
<th>( IG )</th>
<th>( L_{18} )</th>
<th>( LM_{60} )</th>
<th>( LM_{90} )</th>
<th>( \text{Fib.} )</th>
<th>( \text{Unif.} )</th>
<th>( \text{Dic.} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>84.87</td>
<td>84.87</td>
<td>84.87</td>
<td>84.87</td>
<td>84.87</td>
<td>84.87</td>
</tr>
<tr>
<td>20</td>
<td>84.78</td>
<td>85.03</td>
<td>84.69</td>
<td>85.06</td>
<td>85.11</td>
<td>85.07</td>
</tr>
<tr>
<td>40</td>
<td>84.96</td>
<td>85.16</td>
<td>85.10</td>
<td>84.98</td>
<td>85.10</td>
<td>85.05</td>
</tr>
<tr>
<td>60</td>
<td>85.02</td>
<td>85.26</td>
<td>85.24</td>
<td>84.84</td>
<td>85.17</td>
<td>85.01</td>
</tr>
<tr>
<td>80</td>
<td>85.42</td>
<td>85.21</td>
<td>85.28</td>
<td>85.30</td>
<td>85.31</td>
<td>84.81</td>
</tr>
<tr>
<td>100</td>
<td>85.53</td>
<td>85.05</td>
<td>85.72</td>
<td>85.34</td>
<td>85.38</td>
<td>84.98</td>
</tr>
<tr>
<td>120</td>
<td>85.48</td>
<td>84.81</td>
<td>84.87</td>
<td>85.15</td>
<td>85.23</td>
<td>84.79</td>
</tr>
<tr>
<td>140</td>
<td>85.40</td>
<td>84.47</td>
<td>84.96</td>
<td>84.76</td>
<td>84.73</td>
<td>84.85</td>
</tr>
<tr>
<td>160</td>
<td>84.76</td>
<td>83.43</td>
<td>83.73</td>
<td>83.51</td>
<td>83.17</td>
<td>83.84</td>
</tr>
</tbody>
</table>

not even hold, since it is a step function. Hence, only some comparison methods can be applied, like Fibonacci, Golden Search, Uniform Search and Dicotomic Search methods.

3. EXPERIMENTS

The experiments were performed using the Apté split of Reuters-21578 and the split of Ohsumed proposed in \cite{2}.

The methods mentioned above are compared with the best results reported in \cite{1} (\( LM_{60} \) and \( LM_{90} \) for Reuters and \( LM_{18} \) and \( LM_{2} \) for Ohsumed) and in \cite{4} (IG and \( L_{18} \) for Reuters and CET and \( D_{18} \) for Ohsumed). The stop criterion proposed consists in automatically obtain a threshold defined as the minimum difference between the values of \( k = b_{w,c}/a_{w,c} \) of all the pairs \( (a_{w,c}, b_{w,c}) \) \cite{1}. In this way, the target of automatically defining an optimal value of \( k \) is maintained. The evaluation function was \( F_1 \) of a Cross Validation (CV) with 2 folds and 5 repetitions over SVM.

Table 1 and Table 2 show the results \cite{4}. In Reuters-21578, all the methods perform quite similar, being slightly better the Uniform Search for the macroaverage and, in general, the Fibonacci for the microaverage. But, for some filtering levels \( LM_{90} \) performs slightly better for the macroaverage and \( IG \) does it for the microaverage. Performing t-test at the significant level of 95%, we found that there are not appreciable differences. In Ohsumed, the Uniform Search is considerably better than the rest, in fact, the t-test confirms it. Additionally, only \( L_{18} \) is better than the measure obtained by the Uniform Search method for some filtering levels, but in any case, both perform similarly, as t-test stated.

The different behaviour of the methods and of the macroaverage and microaverage over both corpora may be respectively because the distribution of documents over the categories is much more balanced and the words are quite more scattered in Ohsumed than in Reuters-21578.

4. CONCLUSIONS AND FUTURE WORK

This paper proposes an automatic method to obtain optimal Linear Measures for FS in TC. The results over two well-known corpora reveal that the more scattered distribution of words the best the performance of Uniform Search method is. Besides, the goal of automatically selecting a LM is successfully achieved.

The above methods assume that the target function is unimodal, and we are not able to assure it. Hence, the risk of finding a local maximum is latent. Therefore, as future work, we plan to study some alternatives to automatically avoid the local maximums, like genetic algorithms or simulated annealing.

5. ADDITIONAL AUTHORS

Additional authors: Irene Díaz (Artificial Intelligence Center, University of Oviedo, email: sirene@aic.uniovi.es).

6. REFERENCES


