Pushing “Underfitting” to the limit: Learning in Bidimensional Text Categorization

Giorgio Maria Di Nunzio[1] and **Alessandro Micarelli**[2]

Abstract. The analysis of two heuristic supervised learning algorithms for text categorization in two dimensions is presented here. The graphical properties of the bidimensional representation allows one to tailor a geometrical heuristic approach in order to exploit the peculiar distribution of text documents. In particular, we want to investigate the theoretical linear cost of the algorithms and try to push the performance to the limit. The experiments on Reuters-21578 standard benchmark confirm that this approach is an alternative to the standard linear learning models, such as support vector machines, for text classification. Moreover, due to the fast training session, this approach may also be considered as a support for text categorization systems for fast graphical investigations of large collections of documents.

1 Introduction

The *vector space model* is one of the most used models in Information Retrieval for representing documents into the term vector space. Documents are represented with a *term vector*, where each component represents the corresponding dimension of the space. Automated Text Categorization (ATC) usually considers this model [9, 8]. Processing is extremely costly in computational terms by means of standard machine learning techniques since the dimensionality of the space easily reaches hundreds of thousands. Hence, a reduction of the original space is needed. Recent works with different statistical methods, such as Naïve Bayes [2], ridge logistic regression [11], and support vector machines [4], show that it is possible to achieve a good trade-off between efficiency and effectiveness through a feature selection approach [10, 3]. A different approach to the problem is given by projection based methods like: multidimensional scaling [6], and self-organizing maps [5]. These methods serve mainly for exploratory tasks by means of visualization maps that present the overall similarity structure of a corpus of documents.

In this paper we explore the technique in order to project documents into a two-dimensional space presented in [1]. This novel approach makes the compact representation of documents possible, as well as the reduction of complexity from a N-dimensional space to a 2-dimensional one. In addition, a graphical visualization on a 2-D plot of the documents of a collection may be used to analyze the distribution of categories. Our work particularly focuses on the optimization of a supervised learning algorithm that exploits the two-dimensional distribution of documents. We compare the original formulation of the heuristic algorithm, here named *Angular Region*, and a possible reformulation named *Focused Angular Region*. They both share the same hypothesis that, given a cost function, an optimal separating straight line exists between the sets of positive and negative documents. This study aims to find the minimum value of the free-parameters of both algorithms. The experimental results on the standard Reuters-21578 benchmark confirms the linear computational training cost for both algorithms, outperforming the support vector machines in terms of training time, and confirming the two-dimensional approach a valid alternative.

The remainder of the paper is organized as follows: Section 2 presents the definitions of Two-dimensional Text Categorization; in Sect. 3 the two heuristic algorithms are presented; in Sect. 4 the experimental results are shown. The final remarks are given in Section. 5.

2 Two-dimensional Text Categorization

Text Categorization may be formalized as the task of approximating the unknown *target* function $\Phi : D \times C \rightarrow \{T, F\}$ by means of a function $\hat{\Phi} : D \times C \rightarrow \{T, F\}$ called *classifier*, where C is a predefined set of categories and D is the set of documents. Being c_i a generic category $c_i \in C$, for every $d \in D$, if $\Phi(d, c_i) = T$, then d is called *positive example* of c_i, while $\Phi(d, c_i) = F$ it is called *negative example* of c_i (see [8]). ATC is the activity of automatically building automatic text classifiers by means of machine learning techniques. According to the supervised learning approach, an initial corpus of pre-classified documents under some predefined categories is assumed, hereafter called Ω. This corpus is split into two subsets: the *training* set *Tr* and the *test* set *Te*, where *Te* = Ω - *Tr*. The whole *Tr* is used in the experimentation to calculate the statistics of the collection while the heuristic learning algorithms are trained according to the *k-fold cross validation* approach.

In order to have a coherent symbolism among the formulae, some general definitions are given here: we assume to have a set of predefined categories *C* = $\{c_1$, ..., c_i, ..., $c_n\}$, and indicate with $d_{j,i}$ the j-th document that belongs to the i-th category (for example, $c_1 = \{d_{1,1}, d_{j,1}, ..., d_{N_i,1}\}$), each category having a generic number of documents N_i. The set of distinct terms of a category c_i is T_i, while a generic term of the *vocabulary* is $t \in \bigcup_{i=1}^{n} T_i$. Moreover, we indicate with $c_{i|t} \subseteq c_i$ a subset of category c_i, whose elements are the documents of c_i in which the term t appears at least once (for example, $c_{1|t} = \{d_{2,1}, d_{5,1}, d_{11,1}\}$). The cardinality of this subset is indicated by $N_{i|t}$. Finally, the notation c_i indicates the category under investigation, and *RotW* indicates the *rest of the world* which is the difference set $C - c_i$. The words *word*, *term* and *feature* are synonyms.

[1] Dept. of Information Engineering, University of Padua, Padova, Italy. Email: dinunzio@dei.unipd.it

[2] Dept. of Computer Science and Automation, University of “RomaTre”, Roma, Italy. Email: micarel@dia.uniroma3.it

2.1 Document Representation

The skeleton idea on which the body of the whole work lies is as follows: given a set of categories, a generic word may give two different meanings. One is its importance in a particular category (the category of interest), the other is its importance in the other categories (the rest of the world). We often use the terms *local* or *global* in accordance to the aspect to focus. The projection of documents into the bidimensional space requires a supervised learning criterion that starts with the estimate of two parameters: *Presence* and *Expressiveness*. The underlying naïve assumption for the weighting scheme defined here is: the more a term t appears in the set of documents under investigation (c_i or *RotW*) and does not appear in the rest of the collection, the higher the importance of the term for this particular set.

Presence and Expressiveness

Given a category of interest c_i the *local Presence* estimates the relative frequency of the documents which contain a particular term t with respect to the total number of documents of c_i. It is denoted as $\hat{P}(t, c_i)$. The *global Presence* of a term t, denoted as $\hat{P}(t, C - c_i)$, is defined in the following way: for each category c_j belonging to the *RotW*, compute the local Presence $\hat{P}(t, c_j)$ and calculate the arithmetic mean.

Expressiveness exploits the information given by the local and global Presence in an inverted way. In particular, the *local Expressiveness* $\hat{E}(t, c_i)$ is defined as one minus the global Presence. In this way it measures how much a term t, given a category c_i, does *not* appear in the *RotW*. Viceversa, the *global Expressiveness* $\hat{E}(t, C - c_i)$ is defined as one minus the local Presence. It estimates how much the same term does *not* appear in c_i. Table 1 summarizes the definitions given above.

Table 1. Definition of *Presence* $\hat{P}$ and *Expressiveness* $\hat{E}$.

Local	Global
$\hat{P}(t, c_i) = \frac{N_{i\|t}}{N_i}$	$\hat{P}(t, C - c_i) = \frac{\sum_{j=1, j \neq i}^{n} \frac{N_{j\|t}}{N_j}}{n - 1}$
$\hat{E}(t, c_i) = 1 - \hat{P}(t, C - c_i)$	$\hat{E}(t, C - c_i) = 1 - \hat{P}(t, c_i)$

Both measures range from 0 to 1 and match the numerical value with the meaning of the words “presence” and “expressiveness”: the more a term appears (is “present”) in the documents of a category (or in the rest of the world) the higher the Presence; the more a term t is representative (is “expressive”) of a particular category (or for the *RotW*) the higher the value of Expressiveness.

Local and Global Term Weighting

The problem of how to weight a word is seen from two points of view: how to compute the *local importance* of a term in a category c_i with respect to the *RotW*, and how to compute the *global importance* of the same term in the $RotW$ with respect to the c_i.

The local weight of a term is defined as the product of the local Presence and Expressiveness:

$$LW(t, c_i) = \hat{P}(t, c_i) \cdot \hat{E}(t, c_i) \; . \tag{1}$$

The global weight of a term may be seen as the dual problem: consider the rest of the world as our category of interest, and the category c_i as the new $RotW$. The weight is then defined as:

$$GW(t, C - c_i) = \hat{P}(t, C - c_i) \cdot \hat{E}(t, C - c_i) \; . \tag{2}$$

Both weights follow the naïve assumption that a term is more important in a category (or in the *RotW*) if it is present and expressive at the same time. Another interpretation may be: the Presence of a term is penalized by a factor proportional to its Expressiveness.

Local and Global Energy of a Category

The local and global energy are two fictitious measures that summarize the contribution of all the terms in the category. Following the same reasoning of the twofold point of view, the definition of the local energy function LE is defined as the sum of all the local weights. Using Eq. (1):

$$LE_i = \sum_t \hat{P}(t, c_i) \cdot \hat{E}(t, c_i) = \sum_t LW(t, c_i) \; . \tag{3}$$

The global energy function GE is defined as the sum of all the global weights. Using Eq. (2):

$$GE_i = \sum_t \hat{P}(t, C - c_i)) \cdot \hat{E}(t, C - c_i)) = \sum_t GW(t, C - c_i) \; . \tag{4}$$

Bidimensional Coordinates

At this point, each term of the vocabulary has two weights represented by the local and global weight; two measures indicate the local and global energy of a category. The final step of the representation of documents we are looking for should answer the following questions: what is the energy of a document d in the category of interest c_i; and what is the energy of a document d in the $RotW$?

Denoting a generic term that appears in a document d as $\dot{t}$, the coordinate $X_i(d)$ of a document which answers the first point is defined as:

$$X_i(d) = \frac{\sum_{\dot{t} \in d} \hat{P}(\dot{t}, c_i) \cdot \hat{E}(\dot{t}, c_i)}{LE_i} \; , \tag{5}$$

where the energy of the document d in the category c_i is computed by the sum $\sum_{\dot{t} \in d} \hat{P}(\dot{t}, c_i) \cdot \hat{E}(\dot{t}, c_i)$. Accordingly, the second point is answered by the Y_i coordinate of document d:

$$Y_i(d) = \frac{\sum_{\dot{t} \in d} \hat{P}(\dot{t}, C - c_i) \cdot \hat{E}(\dot{t}, C - c_i)}{GE_i} \; , \tag{6}$$

where the energy produced by d in the RotW is the sum $\sum_{\dot{t} \in d} \hat{P}(\dot{t}, C - c_i) \cdot \hat{E}(\dot{t}, C - c_i))$.

Both $X_i(d)$ and $Y_i(d)$ are defined from 0, when d does not contain any term of category c_i (or any term of the rest of the world), to 1 when d contains all the terms of the category (or all the terms of the $RotW$). The coordinates X_i and Y_i form a two-dimensional space that is named here as *space of category* c_i. Figure 1 shows the training documents of Reuters-21578 projected into the space of category *acquisitions*. The stars are the documents that belong to the category of interest, the circles are the *RotW* documents while the solid line represents the point of the space where $X_i = Y_i$. The peculiar distribution of positive and negative documents, which looks like a “V” rotated 45° clockwise, is common to all the categories. Positive documents are almost all below the line $Y_i = X_i$ while some of the negative documents are above the same line and some others overlap with the positive ones.

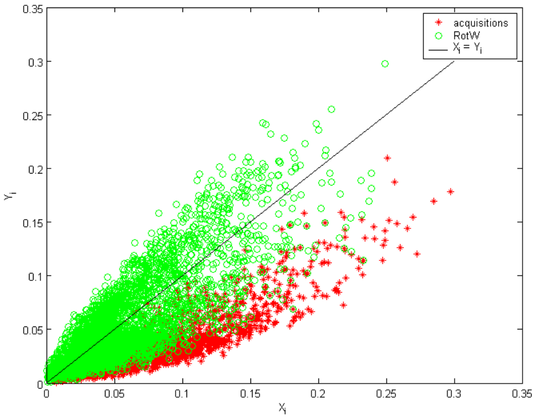

Figure 1. The space of category *acquisition* Reuters-21578

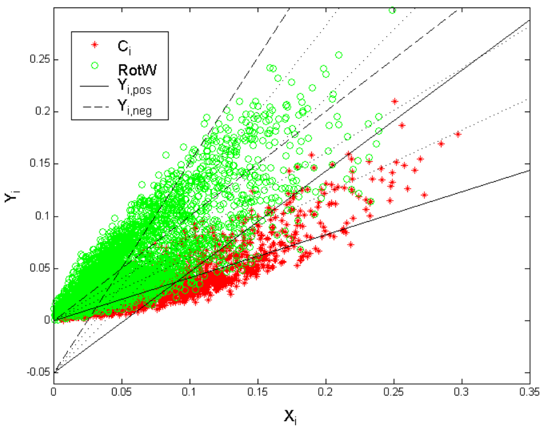

Figure 2. Example of two angular region with $p(0,0)$ and $p(0,-0.05)$

3 Focusing on the Angular Region

Two supervised learning heuristic algorithms are presented here: the *Angular Region* (AR) algorithm and the *Focused Angular Region* (FAR) algorithm. The main hypothesis which lies behind both algorithms is that, given a cost function (in the experiments the F_1 measure defined in Eq. (7)), an optimal separating straight line exists between the sets of positive and negative documents.

The general idea may be stated as follows: let $p(0,q)$ be a point close to the origin, $|q| < 0.1$. Let $Y_{i,pos}$ be the interpolating line of category c_i (positive documents) constrained to pass through the point $p(0,q)$ (note that the interpolating lines of positive/negative documents are found by means of standard vertical least squares fitting procedures):

$$Y_{i,pos} = m_{pos} \cdot X_i + q \ ,$$

and let $Y_{i,neg}$ be the interpolating line of the *RotW* (negative documents) constrained to pass through the point $p(0,q)$

$$Y_{i,neg} = m_{neg} \cdot X_i + q \ .$$

Consider the angular region whose vertex is the point $p(0,q)$, bounded by the semi-lines $Y_{i,pos}$ and $Y_{i,neg}$. Within this region the optimal separating straight line should be found, being the F_1 measure the cost function. The equation of the line would be:

$$Y_{i,opt} = m_{opt} \cdot X_i + q_{opt} \ ,$$

where $m_{pos} \leq m_{opt} \leq m_{neg}$ and $|q_{opt}| < 0.1$. Figure 2 shows two example of angular regions obtained with $p(0,0)$ and $p(0,-0.05)$ using the same training documents of Fig. 1. The dotted lines are some possible optimal separating line with different angular coefficients.

The pseudo-code of AR algorithm is presented in Algorithm 1. The cost of finding the best solution depends on both the resolution of the interval $Q = [q_-, q_+]$ and the resolution of the interval $M = [m_{pos}, m_{neg}]$. We use the notation $|Q|$ and $|M|$, with an abuse of notation, for the resolution of respectively the interval Q and M. Studying the distribution of documents of all the categories of Reuters-21578, we found that the interval Q may be reduced to $[-0.04, 0.0]$.

Algorithm 1 : *Angular Region*

```
Require: $q_-$, $q_+$, $|Q|$, $|M|$, $k$
Return: $m_{opt}$, $q_{opt}$
  for $j = 1$ to $k$ do
    for $q = q_-$ to $q = q_+$ step $\frac{q_+ - q_-}{|Q|}$ do
      compute $m_{pos}$ and $m_{neg}$
      for $m = m_{pos}$ to $m = m_{neg}$ step $\frac{m_{neg} - m_{pos}}{|M|}$ do
        for every $d \in D$ do
          calculate whether $d$ belongs to $c_i$ or to $RotW$
        end for
        calculate the actual $F_1$ measure
      end for
    end for
    store $m_{opt}(j)$ and $q_{opt}(j)$
  end for
  return $m_{opt} = \sum_{j=1}^{k} m_{opt}(j)/k$ and $q_{opt} = \sum_{j=1}^{k} q_{opt}(j)/k$
```

The computational cost for Algorithm 1 is $O(|Q| \times D + |Q| \times |M| \times D)$, where D is the number of documents of the training set. The first part of the cost, $O(|Q| \times D)$, relates to the calculation of the parameters m_{pos} and m_{neg}. When k-fold cross validation is performed the computational cost is multiplied, on a theoretical level, by a factor proportional to the number of subsets k, let us say $O(k \times (|Q| \times D + |Q| \times |M| \times D))$.

During a k-fold cross validation approach, at the j-th iteration the AR algorithm searches the j-th optimal separating line among all the possible values for q and m. Since this line is always close to the final optimal solution, one may think to reduce the size and the resolution of the interval Q and M in order to improve efficiency without degrading performance. For this reason, the FAR algorithm is proposed. At each iteration, this algorithm keeps trace of the average of the parameters q and m of the previous (sub-)optimal solutions in order to “focalize” the space of search; in addition, the reduction of the resolution of Q and M is performed. The pseudo-code of the algorithm is presented in Algorithm 2

4 Experimental Results

The evaluation was carried out on a notebook equipped with an AMD Athlon^{tm} XP 1600+ processor, 256 MB of DDR RAM, on a Mi-

Algorithm 2 : *Focused Angular Region*

```
Require: $q_-$, $q_+$, $|Q|$, $|M|$, $k$
Return: $m_{opt}$, $q_{opt}$
  for $j = 1$ to $k$ do
    if $j > 1$ then
      if $|Q| > 4$ then $|Q|$ = round($|Q|/2$)
      if $|M| > 4$ then $|M|$ = round($|M|/2$)
      $m_{aver} = \sum_{i=1}^{j-1} m_{opt}(i)/(j-1)$
      $q_{aver} = \sum_{i=1}^{j-1} q_{opt}(i)/(j-1)$
      $m_{pos} = m_{aver} - 0.2$ ; $m_{neg} = m_{aver} + 0.2$
      $q_- = m_{aver} - 0.01$ ; $q_+ = m_{aver} + 0.01$
    end if
    for $q = q_-$ to $q = q_+$ step $\frac{q_+ - q_-}{|Q|}$ do
      compute $m_{pos}$ and $m_{neg}$
      for $m = m_{pos}$ to $m = m_{neg}$ step $\frac{m_{neg} - m_{pos}}{|M|}$ do
        for every $d \in D$ do
          calculate whether $d$ belongs to $c_i$ or to $RotW$
        end for
        calculate the actual $F_1$ measure
      end for
    end for
    store $m_{opt}(j)$ and $q_{opt}(j)$
  end for
  return $m_{opt} = \sum_{j=1}^{k} m_{opt}(j)/k$ and $q_{opt} = \sum_{j=1}^{k} q_{opt}(j)/k$
```

crosoft Windows XP Professional OS with Service Pack 1. The algorithms have been implemented in Matlab code (version 6.5 release 13).

The Reuters-21578[3] corpus was chosen as a benchmark. The top 10 most frequent categories of ModApte split were used for experimentation such that the training set was composed of 7193 documents and the test set of 2787 documents. Some text preprocessing was done: a first cleaning was done removing all the punctuation marks and all the numbers and converting all the letters to lowercase. A stoplist of 232 words and contractions (that is, ’re, don’t, etc.) was used to remove the most frequent words of the English language. Finally the English Porter stemmer[4] was used as the only method to reduce the space of terms.

Standard IR evaluation measures have been computed. Recall ρ_i and Precision π_i are defined for each category c_i as (using the same notation of [8]):

$$\rho_i = \frac{TP_i}{TP_i + FN_i} \ , \qquad \pi_i = \frac{TP_i}{TP_i + FP_i} \ ,$$

where TP_i (*true positive*) is the number of documents correctly classified under category c_i, and FN_i (*false negative*) and FP_i (*false positive*) are defined accordingly. The performance of the classifier for the whole set of categories was estimated according to the the F_β function, and in particular when $\beta = 1$:

$$F_\beta = \frac{(\beta^2+1)\cdot\pi\cdot\rho}{\beta^2\cdot\pi+\rho} \ , \qquad F_1 = \frac{2\cdot\pi\cdot\rho}{\pi+\rho} \ . \qquad (7)$$

Then, both the two methods of *micro-averaging* and *macro-averaging* were used to average the performances (see [8]).

The SVM^{Light} package[5], a widely used implementation of SVM, was employed with the default parameters (linear kernel) to compare performances.

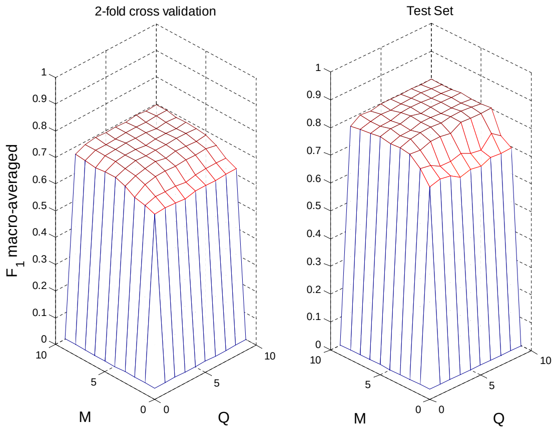

Figure 3. Performance of Algorithm 1 on the validation and test sets

Analysis of the Results

The first test run shed light on the word “underfitting” of the title. We performed a 2-fold cross validation approach for Algorithm 1 using different resolution of Q and M from 2 to 10. The results, summarized in Fig. 3, show that the performance on both the validation and the test set becomes stable, with minor oscillations, when $|Q| > 4$ and $|M| > 4$. This means that it is not necessary to have a high resolution of the two intervals in order to increase the classifier performance. Then we performed 12 runs for each algorithm, varying the number k of subsets from 2 to 5, and using the same resolution of 5, 10 and 15 for Q and M. Each run was repeated ten times using different categories for a total of 240 training sessions.

Figure 4 shows the averaged performance on the validation sets of the Algorithms 1 and 2 (upper graphic) and the averaged training time per category (lower graphic). The combination $|Q| = 5$, $|M| = 5$, $k = 5$, with the FAR algorithm (dash-dotted line) gives the best trade-off between performance and training time. Figure 5 shows the macro- and micro-averaged F_1 performances. Once again, the FAR algorithm with the parameters stated above presents the best results considering both the two measures. Table 2 compares the averaged performance of the 10 training session of the FAR algorithm with respect to the SVM^{Light} on the bidimensional space (a biased hyperplane was used in this case), and with respect to SVM^{Light} on the n-dimensional space using a *TfIdf* weighting scheme (see [7]), without feature selection. The average training time in seconds per category are shown as well.

The results are satisfactory and encourage us to investigate the bidimensional space more accurately. The positive aspects are that both of the heuristic algorithms outperform SVM in terms of training time. Moreover, since SVM are known to optimize the *accuracy* on a given dataset, which is the number of documents correctly classified over the total number of documents, a number of runs for parameter optimization are needed in the two-dimensional space. A drawback is that the average performance is a little bit lower than the state-of-the-art n-dimensional SVM approach. Nevertheless, a category-by-category investigation shows that the performance of the FAR is comparable, if not superior, to the state-of-the-art on seven out of ten categories. Only three categories, *acq*, *wheat* and *grain*, present a bad performance which degrades the overall effectiveness. This peculiar behavior suggests that it may be possible to further improve the performances with a better understanding of the local and global

[3] `http://www.daviddlewis.com/resources/testcollections/reuters21578/`

[4] `http://www.tartarus.org/~martin/PorterStemmer/`

[5] `http://svmlight.joachims.org/`

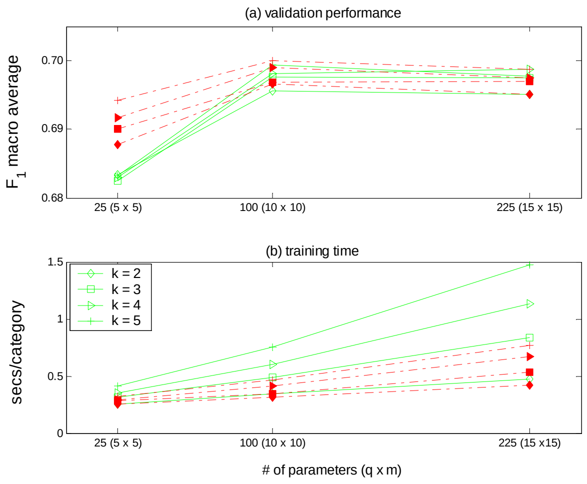

Figure 4. (a) shows the F_1 macro-average of the validation session. Solid lines are the runs with AR algorithm while dash-dotted lines the FAR algorithm. (b) shows the averaged time in seconds to train each category

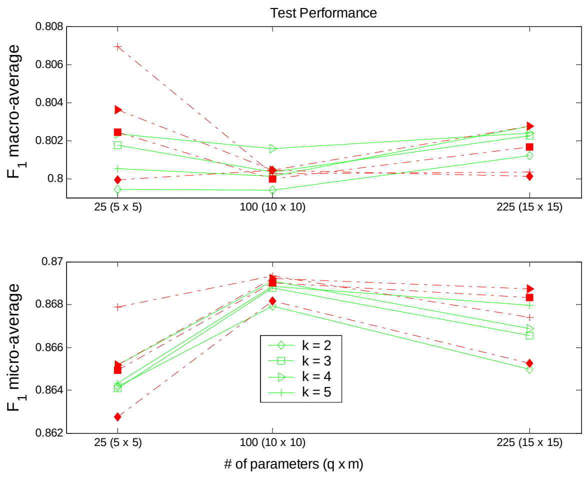

Figure 5. Test Performances on top 10 categories of Reuters-21578. Solid lines are AR algorithm while dash-dotted lines FAR algorithm

parameters.

5 Conclusions

Two heuristic supervised learning algorithms have been investigated in this paper using a new bidimensional representation of documents. These algorithms use the geometrical distribution of documents in order to reduce the search for the optimal separating line. The first algorithm, *Angular Region*, performs a complete investigation in the range of the two parameters (angular coefficient and intercept). The second algorithm, *Focused Angular Region*, is optimized for a k-fold cross validation; at each iteration it uses the sub-optimal parameters given by the previous iteration in order to focalize the range of parameters in a more limited interval. The trade-off between efficiency and effectiveness obtained with these solutions opens new perspectives for fast graphical investigations of large collections of documents for text categorization.

Table 2. F_1 performance comparison among the bidimensional representation and the state-of-the-art. The top 10 categories of Reuters-21578 have been used as benchmark

	Bidimensional Space		N-dimensional space
	FAR	SVM^{Light}	SVM^{Light}(TfIdf)
earn	.946	.911	.986
acq	.855	.830	.962
money-fx	.759	.650	.742
grain	.889	.522	.920
crude	.824	.719	.904
trade	.797	.672	.852
interest	.750	.603	.725
wheat	.784	.380	.828
ship	.845	.601	.844
corn	.596	.311	.867
F_1 Macro	.807	.620	.866
F_1 micro	.868	.777	.929
av. secs/cat.	0.37	2.20	4.07

ACKNOWLEDGEMENTS

The authors would like to give the most sincere thanks to the anonymous reviewers for their helpful comments.

REFERENCES

[1] Giorgio M. Di Nunzio, ‘A bidimensional view of documents for text categorisation’, in *Proceedings of the 26-th European Conference on Information Retrieval (ECIR–04)*, number 2997 in Lecture Notes in Computer Science, pp. 112–126, Sunderland, UK, (April 2004).

[2] Susana Eyheramendy, Daivd D. Lewis, and David Madigan, ‘On the naive bayes model for text categorization’, in *Proceedings of the 9-th International Workshop Artificial Intelligence and Statistics (AISTATS–03)*, Key West, Florida, US, (January 2003).

[3] George Forman, ‘An extensive empirical study of feature selection metrics for text classification’, *Journal of Machine Learning Research*, **3**, 1289–1305, (March 2003).

[4] Thorsten Joachims, ‘Text categorization with support vector machines: Learning with many relevant features’, in *Proceedings of the 10-th European Conference on Machine Learning (ECML–98)*, pp. 137–142, Chemnitz, DE, (1998).

[5] Teuvo Kohonen, *Self-Organizing Maps*, Springer Verlag, Berlin and Heidelberg, DE, 1995.

[6] Joseph B. Kruskal and Myron Wish, *Multidimensional Scaling*, Sage University Paper Series on Quantitative Applications in the Social Sciences, Sage Publications, London, UK, 1978.

[7] Gerard Salton and Chris Buckley, ‘Term-weighting approaches in automatic text retrieval’, *Information Processing & Management*, **24**(5), 513–523, (1988).

[8] Fabrizio Sebastiani, ‘Machine learning in automated text categorization’, *ACM Computing Surveys*, **34**(1), 1–47, (2002).

[9] Yiming Yang, ‘An evaluation of statistical approaches to text categorization’, *Information Retrieval*, **1**(1/2), 69–90, (1999).

[10] Yiming Yang and Jan O. Pedersen, ‘A comparative study on feature selection in text categorization’, in *Proceedings of the 14-th International Conference on Machine Learning (ICML–97)*, pp. 412–420, Nashville, Tennessee, US, (1997).

[11] Thong Zang and Frank J. Oles, ‘Text categorization based on regularized linear classification methods’, *Information Retrieval*, **4**(1), 5–31, (2001).