Angular Measures for Feature Selection in Text Categorization

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

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

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

ABSTRACT
Text Categorization, which consists of automatically assigning documents to a set of categories, usually involves the management of a huge number of features. Most of them are irrelevant or introduce noise which misleads the classifiers. Thus, feature reduction is often performed in order to increase the efficiency and effectiveness of the classification. In this paper we propose to select relevant features by means of what we call Angular Measures, which are simpler than other usual measures applied for this purpose. We carry out experiments over two different corpora and find that the proposed measures perform equal or better than some of the existing ones.

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
One of the main tasks in the processing of large collections of text files is that of assigning the documents of a corpus to a set of previously fixed categories, what is known as Text Categorization (TC) [11]. The most common way of representing the documents for TC is the bag of words (see [10]). In this representation, a vector is associated to each document whose components quantify the importance of each of its words. This usually involves a great amount of features and most of them can be irrelevant or noisy [10]. Thus, feature reduction often leads to an improvement in the performance of the classification, at the same time that it reduces the computational cost and the storage requirements of the task.

A common approach for feature reduction is Feature Selection (FS), which consists in choosing a subset of the original features for representing the documents. This task in TC is usually performed scoring the features using a certain measure, ordering them according to this measure and removing a predefined number or percentage of them [8, 13]. Several measures have been proposed for this purpose, like information gain [13] or cross entropy for text [8].

In this paper we introduce some measures for FS in TC, which we call Angular Measures. We define them and study their behavior by means of experimentation over two well known corpora.

The paper is organized as follows. Section 2 deals with some previous work including some of the state-of-the-art measures. Section 3 presents the new family of measures proposed. Section 4 describes the main stages of the TC task. The description of the corpora and the experiments are detailed in Sections 5 and 6 respectively. Finally, in Section 7 some conclusions and ideas for future work are commented.

2. PREVIOUS WORK
FS is one of the approaches commonly adopted in TC. It involves selecting a subset of features from the original feature set. By contrast, Feature Extraction (FE) methods transform or combine the original features to obtain a reduced number of features. Methods of this kind are clustering ones [4] or Latent Semantic Indexing (LSI) [6].

On the other hand, John et al. distinguish two kinds of FS, namely filtering and wrapping. In the former, a feature subset is selected independently of the performance of the classifier. In the latter, a feature subset is selected using an evaluation function based on the classifier. A widely adopted approach in TC is the filtering one based on selecting the features with higher score granted by a certain measure. The reason of preferring filtering approaches rather than wrappers for TC is that the latter usually result in a considerably time consuming process.

In the following paragraphs will briefly describe those measures which have been most adopted for FS in TC.

2.1 Statistical Measures
The simplest filtering measures are the term frequency (tf) and the document frequency (df). They quantify the relev-
vance of a word by means of its total number of appearances and by means of the number of different documents in which it appears, respectively. They can be combined into \( tfidf \) \[^{[10]}\] defined by

\[
tfidf = tf \log \frac{N}{df}
\]

where \( N \) is the number of documents in the corpus. Notice that words appearing in all the documents are considered non-informative, independently of its absolute frequency and, in general, a word occurring in many documents will have \( tfidf \) smaller than others with the same \( tf \), but appearing in less documents. Despite their simple appearance, these measures perform acceptably in many situations \[^{[5]}\].

### 2.2 Information Theory Measures

Measures taken from Information Theory (IT) have been widely used because it is interesting to consider the distribution of a word over the different categories. Among these measures, information gain (IG) takes into account the presence of the word in a category as well as its absence, and can be defined by (see, for instance \[^{[13]}\])

\[
IG(w, c) = P(w)P(c|w) \log \frac{P(c|w)}{P(c)} + P(w\bar{c})P(c|\bar{w}) \log \frac{P(c|\bar{w})}{P(c)}
\]

where \( P(w) \) is the probability that the word \( w \) appears in a document, \( P(c|w) \) is the probability that a document belongs to the category \( c \) knowing that the word \( w \) appears in it, \( P(c|\bar{w}) \) is the probability that the word \( w \) does not appear in a document and \( P(c|\bar{w}) \) is the probability that a document belongs to the category \( c \) if we know that the word \( w \) does not occur in it. Usually, these probabilities are estimated by means of the corresponding relative frequencies.

Another measure of this kind is the expected cross entropy for text (CET) \[^{[8]}\], which only takes into account the presence of the word in a category. It is defined by

\[
CET(w, c) = P(w)P(c|w) \log \frac{P(c|w)}{P(c)}
\]

These measures are the ones of this kind that have obtained better results in TC \[^{[7, 8, 11, 13]}\].

### 3. Angular Measures

Before defining this family of measures, let consider a category \( c \) and a word \( w \) and identify each word \( w \) with the pair \((a_w, b_w)\), where \( a_w \) denotes the number of documents of the category \( c \) in which \( w \) appears and \( b_w \) denotes the number of documents that contain the word \( w \) but do not belong to the category \( c \). On what follows let denote the pair \((a_w, b_w)\) by \((a, b)\) for simplicity.

Then, we can study the words that receive identical score under a filtering measure \( m(w) \) which depends only on \( (a, b) \) by means of the level curves defined by such measure. In fact, it was demonstrated in \[^{[2]}\] that if \( m(w) \) is a filtering measure and \( N \) and \( M \) are natural numbers, then the level curves passing through the words with \( a_w \leq N \) and \( b_w \leq M \) can be considered as straight lines.

From that fact, an interesting special case is the family of measures \( m(w) \) which have just one level curve for each value. That is, the measures \( m(w) \) that satisfy

\[
a_w = f(m(w))b_w + g(m(w))
\]

for some functions \( f \) and \( g \). Some measures \[^{[2]}\] like \( df \) (with \( f(df) = -1 \) and \( g(df) = df \)), have this property.

In \[^{[2]}\] it has also been proven that if \( N \) and \( M \) are two natural numbers and \( m(w) \) is a filtering measure which has exactly one straight line as level curve for each value that \( m(w) \) attains over the words with \( a_w \leq N \) and \( b_w \leq M \), then, there exist \( p \) and \( q \) two polynomials such that

\[
a_w = p(m(w))b_w + q(m(w))
\]

for any word \( w \) such that \( a_w \leq N \) and \( b_w \leq M \).

Therefore, it is interesting to study the filtering measures satisfying the above expression at least when the degree of \( p \) and \( q \) is low. The family of measures obtained when \( degree(p) = 0 \) and \( degree(q) = 1 \) has been studied in \[^{[2]}\], leading to what we call Linear Measures. This paper deals with those ones obtained when \( degree(p) = 1 \) and \( degree(q) = 0 \) (notice that it is not possible that the degrees of \( p \) and \( q \) are both zero at the same time).

Thus, if \( degree(p) = 1 \) and \( degree(q) = 0 \) we have

\[
a_w = (c_1m(w) + c_2)b_w + c_3
\]

for some constants \( c_1, c_2 \) and \( c_3 \) such that \( c_1 \neq 0 \), and thus

\[
m(w) = \frac{a_w - c_3}{b_w} - c_2
\]

or equivalently

\[
m(w) = \frac{a_w - c_2b_w - c_3}{c_1b_w}
\]

The value of \( c_1 \) can be taken as 1 since it does not affect the ordering of words produced by the measure. Then, we obtain

\[
m(w) = \frac{a_w - c_2b_w - c_3}{b_w}
\]

with \( c_2 \) and \( c_3 \) any real numbers. But the above expression is equivalent to the following one

\[
m(w) = \frac{a_w - c_3}{b_w} - c_2
\]

and, again, the value of \( c_2 \) is irrelevant in the sense that the ordering of the words provided by the measure is independent of the value of this constant. Hence, the value of \( c_2 \) could be taken to be zero. Therefore, the family of measures to study are of the form

\[
m(w) = \frac{a_w - c_3}{b_w}
\]

or equivalently

\[
AM_k(w) = \frac{a_w - k}{b_w}
\]

where \( k \) is a real parameter which defines the family. These measures have a simple geometrical interpretation as the next Theorem establishes and whose proof can be found in \[^{[9]}\].

**Theorem 1.** The value \( AM_k(w) \) is the tangent of the angle formed by the \( x \)-axis and the line determined by the points \((a_w, b_w)\) and \((k, 0)\).

This is the reason why we will call these measures Angular Measures.
4. TASK OF TEXT CATEGORIZATION

This section describes the stages of the TC task.

The bag of words [10] model is adopted for representing the documents. It consists in viewing a document as a set of words without order and structure. Also, tf is chosen to quantify the importance of each word in each document, since it is one of the most used in the literature [8, 11].

The classification stage consists in assigning a category to a document from a finite set of m categories. This is commonly converted into m binary problems, each one consisting of determining whether a document belongs to a fixed category or not. This approach is called one-against-the-rest [1].

That process leads to use different sets of words in the document representation. One consists of words that belong to each category isolated from the rest, which is known as local approach. On the other hand, the global approach considers the words from all categories. In this work, the local approach is considered, since they offer better results [11].

Additionally, the stop words (words without meaning) are removed because they are useless for the classification. Also, stemming is performed, which consists in mapping the words with the same meaning but with slight different spelling into a common root. The Porter algorithm [9] is adopted for this purpose.

In this paper the classification is performed using Support Vector Machines (SVM) [7], since they have shown to perform fast and well in TC [12]. They satisfactorily deal with many features and with sparse examples. They are binary classifiers which find out threshold functions to separate the documents of a certain category from the rest. We adopt a linear threshold since most TC problems are linearly separable [7].

The popular and well known measure $F_1$ [11] is adopted in this paper to evaluate the effectiveness of the TC task. It is defined by

$$F_1 = \frac{1}{0.5 \frac{P}{R} + 0.5 \frac{R}{P}}$$

where P quantifies the percentage of documents that are correctly classified as belonging to the category while R quantifies the percentage of documents of the category that are correctly classified.

To compute the global performance over all the categories, macroaverage, which consists in averaging the values obtained in each category [11], is used.

5. THE CORPORA

In this subsection the corpora used in the experiments are described and analyzed. They are the Reuters-21578 collection and the Ohsumed collection.

5.1 Reuters-21578 Collection

The Reuters-21578 corpus is a set of economic news published by Reuters in 1987\(^1\). They are distributed over 135 categories. Each document belongs to one or more of them. The split into train and test documents chosen is that of Apté [1]. Removing some documents without body or topics, 7063 training and 2742 test documents assigned to 90 categories are obtained.

\(^1\)It is publicly available at http://www.research.att.com/lewis/reuters21578.html

The distribution of documents into the categories is quite unbalanced. In fact, the relative dispersion of the number of documents of the categories is 3.36% in the interval [1, 2709] for training documents and 3.39% in [1, 1044] for test documents. In addition, 76.40% (in train) and 78.65% (in test) of the categories have less than 1% of the documents.

The words in the corpus are little scattered, since almost half (49.91%) of the words appear in only one category and 16.25% in only two categories.

5.2 Ohsumed Collection

Ohsumed is a MEDLINE subset of references from 270 medical journals over 1987-1991\(^2\). They are classified into the 15 fixed categories of MeSH\(^3\): A, B, C ... Each category is in turn split into subcategories. We have taken the first 20000 documents of 1991 with abstract, labelling the first 10000 documents as training and the rest as test ones following [7]. We split them into the 23 subcategories of category C of MeSH again following [7].

The distribution of documents over the categories is much more balanced than in Reuters. In fact, the relative dispersion of the number of documents of the categories is 0.86 in the interval [100, 2476] for train and 0.88% in the interval [82, 2424] for test. Furthermore, only 4.35% in train and 8.70% in test of the categories have less than 1% of the documents, against about 77% in Reuters.

The words in this collection are quite more scattered than in Reuters, since 19.55% of the words (in average) appear just in one category (against 49.91% in Reuters).

6. THE EXPERIMENTS

In the theoretical study developed in [3] we have proved that the values of k of the form

$$\frac{a_w b_v - a_v b_w}{b_w - b_v}$$

with v, w two words of the collection are relevant, since they provide measures which discriminate words from one category from the rest. Hence, due to this fact and as a first approach we select the deciles of the distribution formed by those values (when w ranges over all the words of the category under study) as candidates values of k.

Figures 1, 2, 3 and 4 show the macroaverage of $F_1$ of those deciles for Reuters and Ohsumed respectively. Also, they show a comparison of them with two well know and good IT measures, CET and IG, as mentioned in Section 2.

In both corpora, the value of $F_1$ progressively increases from the 1st decile to the median and it decreases until the 9th decile, being the median an inflection point where the maximum is reached.

In the case of Reuters, only the median beats the state-of-the-art measures CET and IG, meanwhile in Ohsumed all the deciles from the 1st to the median achieve better results.

The different behavior of the Angular Measures in both corpora might be due to the different nature of the collections. As we have already mentioned, the distribution of documents into the categories in Reuters is quite more unbalanced than in Ohsumed. Also, the words in Reuters are little scattered in comparison to Ohsumed.

\(^2\)It can be found at http://trec.nist.gov/data/t9-filtering

\(^3\)Available at www.nlm.nih.gov/mesh/2002/index.html
It is also remarkable that some of the Angular Measures obtain the best performance for very high filtering levels (around 95%), specially in the Ohsumed collection. This makes these measures a very appealing choice when an aggressive reduction of the number of features is intended.

7. CONCLUSIONS AND FUTURE WORK

This paper presents a family of measures called Angular Measures for Feature Selection in Text Categorization. They are obtained from their level curves and are defined by a parameter whose adequate values have been carefully selected. The median of certain distribution strategically chosen offers the best results, beating some of the state-of-the-art measures for the two corpora taken. Also, some deciles of such distribution beat those measures in one of the corpora. Additionally, the best performance is obtained when most of the words are removed (about 95% of them) which allows to conduct aggressive reductions using this family of measures.

As future work, we plan to perform a refinement of the values of the parameter taking into account the centiles around the median of the distribution chosen. We also plan to propose several modifications of the Angular Measures based on the performance of other state-of-the-art measures.

8. ACKNOWLEDGMENTS

The research reported in this paper has been supported in part under MEC and FEDER grant TIN2004-05920.
9. ADDITIONAL AUTHORS
Additional authors: Irene Díaz (Artificial Intelligence Center, University of Oviedo, email: sirene@aic.uniovi.es).

10. REFERENCES