The impact of NLP techniques in the text multilabel classification problem

Teresa Gonçalves and Paulo Quaresma

Departamento de Informática,
Universidade de Évora,
7000 Évora, Portugal
tcg|pq@di.uevora.pt

Abstract. Support Vector Machines have been used successfully to classify text documents into sets of concepts. However, typically, linguistic information is not being used in the classification process or its use has not been fully evaluated.

In this paper we apply and evaluate two basic linguistic procedures (stop-word removal and stemming/lemmatization) to the multilabel classification problem of text documents.

These procedures are applied to the Reuters dataset[1] and to the Portuguese juridical documents from Supreme Courts and Attorney General's Office. The obtained results are presented and evaluated.

1 Introduction

Automatic classification of documents is an important problem in many domains. For instance, it is needed by web search engines and information retrieval systems in order to organise text bases into sets of semantic categories.

In order to develop better algorithms for document classification it is necessary to integrate research from several areas, such as, machine learning, natural language processing, and information retrieval.

In this paper we evaluate the use of natural language processing (NLP) and information retrieval (IR) techniques to improve machine learning algorithms, namely:

- IR techniques – stop words removal, documents as bag-of-words;
- NLP techniques – stemming or lemmatization;
- Machine learning algorithm – Support Vector Machines.

Since the work of Joachims [4] it is known that Support Vector Machines (SVM) perform quite well compared with other approaches to the text classification problem. In his approach, documents are represented as bag-of-words (without word order information) [7] and in some experiments, some words are not represented (words belonging to the set of the so called "stop-words").

[1] Available at http://www.research.att.com/l̃ewis/reuters21578.html

A kernel based learning algorithm is then applied (SVM [2]) and the results are evaluated using information retrieval measures, such as the precision-recall break-even point (PRBP).

In this paper, we follow Joachims approach, having as major goal the evaluation of using linguistic information in the classification problem.

We have chosen two sets of documents written in two different languages (English and Portuguese) – the Reuters dataset and the Portuguese Attorney General's Office dataset (PAGOD) [6] – and evaluated the impact of using stop-word removal and stemming/lemmatization in the multilabel classification problem.

In section 2 our classification problem is described and characterised. In section 3 a brief description of the Support Vector Machines theory is presented. Section 4 describes our experiments and evaluates the results and in section 5 some conclusions and future work are pointed out.

2 Text Classification

Our goal is to evaluate the use of linguistic information in the multilabel classification problem of documents, i.e. documents can be classified into multiple concepts/topics.

The typical approach to the multilabel classification problem is to divide it into a set of binary classification problems, where each concept is considered independently. In this way, the initial problem is reduced to solve several binary classification problems.

Binary classification problems can be characterised by the inference of a classification rule assigning one of two possible values $(-1, 1)$ to each document. A value of -1 means the document does not belong to the concept and a value of 1 means it belongs to it.

2.1 Reuters dataset

The Reuters-21578 dataset was compiled by David Lewis and originally collected by the Carnegie group from the Reuters newswire in 1987. We used the "ModApte" split, that led to a corpus of 9603 training documents and 3299 test documents. There are 135 potential categories, but only 90 of those appear at least once both in train and test sets. This collection is known to have a restricted vocabulary: the training set contains only 27658 distinct terms. The classification task is to assign articles to a set of topics. For many of the categories there is a direct correspondence between words and themselves (for example, the appearance of the *wheat* words is a very good predictor of the *wheat* category.)

Next table presents the five top categories and their frequencies for the train and test sets.

category	#train	#test
earn	2861	1080
acq	1648	718
money-fx	534	179
grain	428	148
crude	385	186

2.2 PAGOD dataset

These documents represent the decisions of the Attorney General's Office since 1940 and define a set with cardinality 8151 and around 96MB of characters. All documents were manually classified by juridical experts into a set of classes belonging to a taxonomy of legal concepts with around 6000 terms. However, a preliminary evaluation showed that only around 3000 terms are used in the classification.

Next table presents the top five categories and their frequencies:

id	category	#doc
1391	pensao por servicos excepcionais e relevantes	909
2572	deficiente das forcas armadas	680
744	aposentacao	497
16	funcionario publico	410
204	militar	409

2.3 Document representation

An important open problem is the representation of the documents. In this work, we will use the standard vector representation [7], where each document is represented as a bag-of-words and where order information is lost and no syntactical or semantical information is used.

However, in some experiments, documents were pre-processed in order to remove stop-words and to transform each word in its stem or lemma (for instance, each verb is transformed into its infinitive form and each noun to the singular form).

In the Portuguese set of documents this work was done using a Portuguese lexical database – POLARIS – allowing the lemmatization of every Portuguese word. In the Reuters dataset we used the Porter algorithm [5] to transform each word into its stem.

3 Support Vector Machines

In this section a brief introduction to kernel classifiers and support vector machines is presented[2]. More detailed information can be obtained in several specialised books, such as [8,3].

[2] This introduction is based on a similar section in [1]

Kernel learning algorithms are based on theoretical work on statistical learning theory, namely the structural risk minimisation [10,9].

A binary classifier is a function from an input space X into the set of binary labels $\{-1, +1\}$. A supervised learning algorithm is a function assigning, to each labelled training set, a binary classifier

$$h : X \rightarrow \{-1, +1\} \tag{1}$$

Whenever X is a vector space, a simple binary classifier is given by:

$$h(x) = sign(< w, x > +b) \tag{2}$$

where $< ., . >$ stands for the vector dot-product.

Learning the linear classifier is equivalent to finding values for w and b, which maximise an evaluation measure.

Linear classifiers fail when the boundary between the two classes is not linear. In this situation the approach followed is to project X into a new feature space F and to try to define a linear separation between the two classes in F. If the projection function is defined by $\phi : X \rightarrow F$ then the linear classifier is:

$$h(x) = sign(< w, \phi(x) > +b) \tag{3}$$

Support Vector Machines (SVM) are specific learning algorithms for linear classifiers, trying to obtain values for w and b. In SVM w is assumed to be defined as a linear combination of the projections of the training data:

$$w = \sum_{i=1}^{l} y_i \alpha_i \phi(x_i) \tag{4}$$

where α_i is the weight of the training example i with input x_i and label y_i.

The optimal weights are the solution of a high dimensional quadratic problem, which can be expressed in terms of the dot product of the projection of the training data $< \phi(x_i), \phi(x_j) >$.

It was proved that it is not necessary to map the input data into the feature space F, as long as a kernel function $K : X * X \rightarrow R$, such that $K(x, y) =< \phi(x), \phi(y) >$ is defined. This is known as the *kernel trick*. On the other hand Mercer's theorem [3] states that any positive semi-definite symmetric function corresponds to some mapping in some space and that it is a valid kernel.

In the scope of this work only linear kernels are used and each document is represented by a vector where each dimension value stands for the frequency of a specific word in that document.

4 Experiments

As referred in the previous sections, the SVM learning algorithm was applied to the problem of multilabel classification of the Reuters and PAGOD datasets.

We have chosen the top five concepts of each dataset (the most used ones) and, for each of them, we have performed the following classes of experiments:

- feature selection
- base vs stop-words removal vs stemming

In the first class of experiments we evaluated the impact of the selection of words/features in the performance of the classifier and, in the second, the impact of removing stop-words and performing stemming (or lemmatization, for the Portuguese dataset).

In the base experiments, documents are represented by the set of their words (and frequencies); in the stop-words experiments, words belonging to a specific set (articles, pronouns and prepositions) are removed; in the stemming experiments, stop-words are removed and every word is transformed in its stem (or in its lemma – Portuguese dataset).

For the evaluation we analysed the precision, recall and F_1-*measure*. F_1 belongs to a class of functions used in information retrieval, the F_β-*measure*, that combines the precision and the recall measures [7]. Precision and recall are calculated from the contingency table of the classification (prediction vs manual classification). Precision is given by the number of correct classified documents divided by the number of documents classified to belong to the class. Recall is given by the number of correct classified documents divided by the number of documents belonging to the class.

All the SVM experiments were done using the WEKA software package [11] from Waikato University[3] with default parameters for all experiments).

4.1 Reuters dataset

Feature Selection For the top 5 classes, we have analysed the F_1-*measure* after selecting the number of features/words. The selection was done by eliminating words that appear in less than a specific number of documents: $r55$ means that all words appearing in less than 55 documents were eliminated.

From the analysis of the results it is possible to conclude that concepts in the Reuters dataset have similar behaviour: the performance slightly increases in the beginning and, then, it decreases when the number of the selected words is decreased (figure 1 shows the results for the two top concepts).

A more refined evaluation shows that $Min\_hits = 50$ is a good option for feature selection: it allows the deletion of some non-relevant words and it slightly increases performance.

Linguistic Information Evaluation In this section we tried to evaluate the impact of using linguistic information in the classifier performance. We have chosen one of the best values for Min_hits obtained in the previous

[3] http://www.cs.waikato.ac.nz/ml/weka

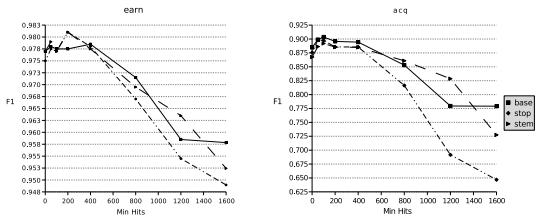

Fig. 1. Reuters feature selection

section (50) and compared the results for the top 5 concepts classification in the following situations:

- base – no use of linguistic information;
- stop-words – removing a list of considered non-relevant words, such as, articles, pronouns, adverbs, and prepositions;
- stemming – removing a list of non-relevant words and transforming each word into its stem.

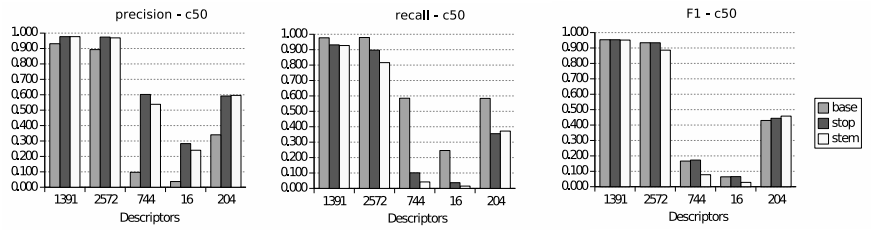

Fig. 2. Reuters linguistic evaluation

Figure 2 shows the obtained results. It is possible to observe that the performance is similar in the three experiments.

This fact allows us to formulate the hypothesis that SVM learning algorithms deal quite well with non-informative features, such as stop-words, and with classes of non-independent features, such as several variations of a specific word.

4.2 PAGOD dataset

Feature Selection We have applied the feature selection methodology described in the previous section to the PAGOD dataset.

The results for the top two concepts are presented in figure 3 (we used a 10-fold cross validation evaluation procedure).

From the analysis of the results for the top concepts it is possible to conclude that, in this case, there is no common behaviour for every concept.

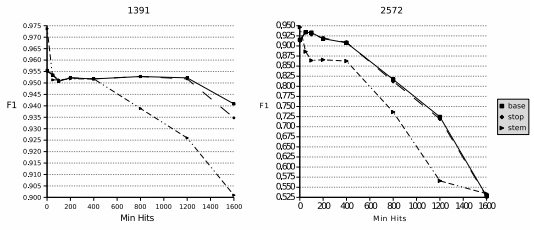

Fig. 3. PAGOD feature selection

This result is quite different from the obtained for the Reuters dataset and it requires further research. In a preliminary approach, we believe these results may be explained by the existence of concepts with distinct levels of abstraction. For instance, we have very specific concepts, such as, "pension for relevant services", but we also have more generic concepts, such as, "public service". The classification of abstract concepts is more difficult and require a more complex approach, such as, the use of semantic information in the classification procedure.

Nevertheless, it is possible to define the value 50 has a good trade-off for the choice of the *Min_hits* variable: it presents the best results for the majority of the concepts.

Linguistic Information Evaluation The PAGOD dataset was also evaluated for the relevance of the *base*, *stop-words* and *stemming/ lemmatization* experiments.

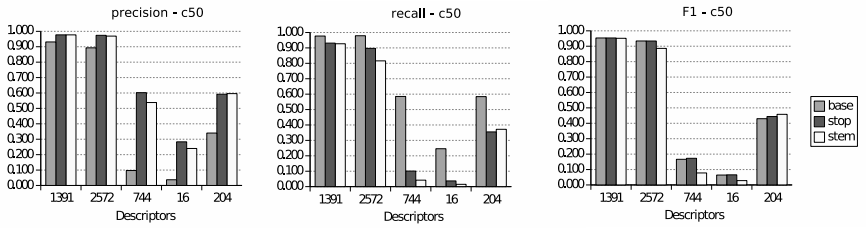

Fig. 4. PAGOD linguistic evaluation

As it can be seen in figure 4 the top 5 concepts show quite distinct results. Some concepts have quite good precision and recall measures, but others have quite bad results. As referred in the previous sub-section, this fact is, probably, related with the existence of concepts with different levels of abstraction and, as a consequence, with different levels of difficulty to be classified. Nevertheless, we can observe that *base*, *stop-words*, and *stemming/ lemmatization* experiments show similar results. These results also support the hypothesis formulated in the Reuters dataset section: SVM learning algorithms deal

quite well with non-informative features and with non-independent features. In fact, these results allow us to extend this hypothesis to non English languages, such as, the Portuguese one.

5 Conclusions and Future Work

The impact of using some natural language processing techniques, such as, lemmatization and removal of non informative words, in the text classification problem was evaluated.

The evaluation was done using the Reuters and PAGOD (in Portuguese) datasets.

The obtained results allowed us to formulate the following hypothesis: SVM learning algorithms deal quite well with non informative and with non independent features in different languages (English and Portuguese).

Nevertheless, for some concepts in the PAGOD dataset, the obtained results were not quite good and further work needs to be done in order to explain them and to improve the classifiers. Our hypothesis is that these classifiers need more powerful document representations, such as the use of word order and syntactical and/or semantical information. This is what we intend to explore as future work.

References

1. N. Cancedda, E. Gaussier, C. Goutte, and J. Renders. Word sequence kernels. *Journal of Machine Learning Research*, 3:1059–1082, 2003.
2. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
3. N. Cristianini and J. Shawe-Taylor. *Support Vector Machines*. Cambridge University Press, 2000.
4. Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines*. Kluwer academic Publishers, 2002.
5. M. Porter. An algorithm for suffix stripping. *Program (Automated Library and Information Systems)*, 14(3):130–137, 1980.
6. Paulo Quaresma and Irene Pimenta Rodrigues. PGR: Portuguese attorney general's office decisions on the web. In Bartenstein, Geske, Hannebauer, and Yoshie, editors, *Web-Knowledge Management and Decision Support*, Lecture Notes in Artificial Intelligence LNCS/LNAI 2543, pages 51–61. Springer-Verlag, 2003.
7. G. Salton and M. McGill. *Introduction to Modern Informatin Retrieval*. McGraw-Hill, 1983.
8. B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.
9. V. Vapnik. *Estimation of Dependencies based on Empirical Data*. Springer, 1982.
10. V. Vapnik. *The nature of statistical learning theory*. Springer, 1995.
11. I. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, 1999.