The impact of NLP techniques in the multilabel text 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.

We apply and evaluate two basic linguistic procedures (stop-word removal and stemming/lemmatization) to the multilabel text classification problem.

These procedures are applied to the Reuters dataset and to the Portuguese juridical documents from Supreme Courts and Attorney General's Office.

1 Introduction

Automatic classification of documents is an important problem in many domains. Just as an example, it is needed by web search engines and information retrieval systems 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. In this paper, we evaluate the impact of using natural language processing (NLP) and information retrieval (IR) techniques in a machine learning algorithm. Support Vector Machines (SVM) was the chosen classification algorithm.

Several approaches to automatically classify text documents using linguistic transformations and Machine Learning have been pursued. For example, there are approaches that present alternative ways of representing documents using the RIPPER algorithm [10], that use Naïve Bayes to select features [5] and that generate of phrasal features using both [2].

In Joachims's [4] approach, documents were represented as bag-of-words (without word order information) [8], a kernel based learning algorithm was applied (SVM [1]) and the results were evaluated using information retrieval measures, such as the precision-recall break-even point (PRBP).

In this paper, we follow Joachims approach, having as a major goal the performance 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 and the Portuguese Attorney General's Office (PAGOD) [7] datasets.

Section 2 presents a brief description of the Support Vector Machines theory. Section 3 characterises our classification problem and section 4 describes our experiments. Section 5 points out some conclusions and future work.

2 Support Vector Machines

Support Vector Machines (SVM) belong to the group of kernel learning algorithms.These algorithms come from the area of statistical learning theory and are based on the structural risk minimisation principle [11].

SVM are supervised linear binary classifiers and, as such, they fail to present a solution when the boundary between the two classes is not linear. In this situation the approach followed is to project the input space X into a new feature space F and try to define a linear separation between the two classes in F. In this way, SVM classifiers can be obtained using algorithms that find the solution of a high dimensional quadratic problem.

In the scope of this work only linear kernels (functions that transform the input feature space). More detailed information can be obtained in several specialised books, such as [9].

3 Text Classification

Our documents classification problem can be characterised as a multilabel one, i.e. documents can be classified into multiple concepts/topics.

The typical approach to solve it, 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.

An important open problem is the representation of the documents. In this work, we will use the standard vector representation [8], where each document is represented as a bag-of-words (retaining words' frequencies). In this representation order information is lost and and no syntactical or semantical information is used.

3.1 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 us to a corpus of 9603 training and 3299 test documents.

The most used category (the top one), *earn*, appears in 2861 training documents and in 1080 test ones, while the fifth most used, *crude*, appears only in 385 and 186 documents, respectively.

3.2 PAGOD dataset

These documents, written in Portuguese, 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.

From the 8151 documents, 909 are classified in the top category, *pensão por serviços excepcionais e relevantes* (*pension for relevant services*), while only 409 documents are classified in the fifth one, *militar* (*military*).

4 Experiments

We chose the top five concepts (the most used ones) from the Reuters and PAGOD datasets and applied the SVM learning algorithm. For each dataset we performed two classes of experiments. In the first 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).

For the evaluation we analysed the accuracy, precision, recall and F_1-*measure*. F_1 belongs to a class of functions used in information retrieval, the F_β-*measure*. This measure combines the precision and the recall ones [8].

All SVM experiments were done using the WEKA software package[12] from Waikato University (with default parameters for all experiments) and the significance tests were made with 95% of confidence level. For the PAGOD dataset we performed a 10-fold cross validation procedure.

4.1 Feature Selection

The feature selection was done by eliminating words that appear less than a specific number in all documents: $s55$ means that all words appearing less than 55 times in all documents were eliminated. We performed experiences for $s1$, $s50$, $s100$, $s200$, $s400$, $s800$, $s1200$ and $s1600$.

Figure 1 shows F_1-*measure* for the *positive* class, for different values of s and the top concept of both datasets.

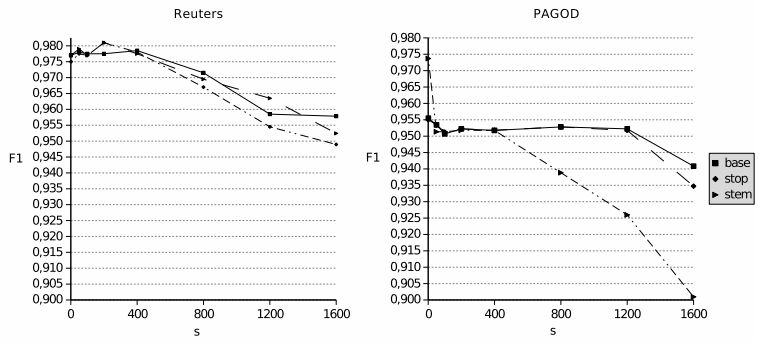

Fig. 1. F_1 for different values of feature selection

Looking at the SVM performance, it is possible to say that $s50$ is a good trade-off for feature selection: it allows the deletion of some non-relevant words and it slightly increases performance.

Analysing the results for the 5 top categories, one can conclude that while, in the Reuters dataset all studied concepts have similar behaviour (the performance slightly increases in the beginning and, then, it decreases when the number of the selected words is decreased), in the PAGOD dataset there is no common behaviour between them.

While the experiments for the Reuters dataset were not statistically different, *stem* is better than *base* and *stop* for $s1$, and worst than the others for $s1200$ and $s1600$, for the PAGOD dataset.

4.2 Linguistic Information Evaluation

For $s50$, we made the following experiments:

- *base* – no use of linguistic information;
- *stop* – removing a list of considered non-relevant words, such as, articles, pronouns, adverbs, and prepositions;
- *stem* – removing a list of non-relevant words and transforming each word into its stem (or to its lemma for the Portuguese).

In the PAGOD dataset this work was done using a Portuguese lexical database, POLARIS, that allows the lemmatization of every Portuguese word. In the Reuters dataset we used the Porter algorithm [6] to transform each word into its stem.

Figure 2 shows F_1-*measure* for the *positive* class, for the 3 experiments and the five top concepts of both datasets.

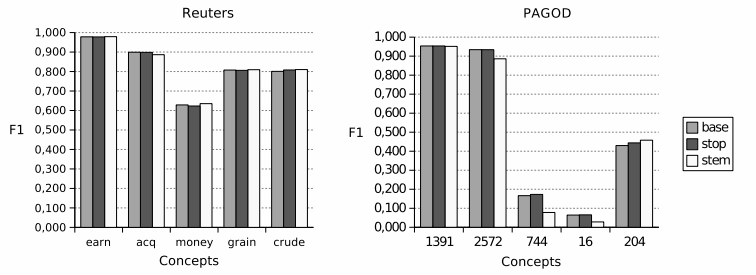

Fig. 2. F_1 for *base*, *stop-words* and *stem* experiments

For both datasets, we can observe that *base*, *stop* and *stem* experiments show similar results. Concerning the 5 categories, the Reuters dataset present similar performances, while the PAGOD dataset doesn't.

The experiments for the Reuters dataset are not statistically different, but, for the PAGOD dataset *stem* is worse than *base* and *stop* for the concepts 2572, 744 and 16.

5 Conclusions and Future Work

The Reuters dataset presents good values for the studied measures and similar results for all studied concepts. On the other hand, for the PAGOD dataset some concepts have quite good precision and recall measures, but others have quite bad results.

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. Our hypothesis is that these classifiers need more powerful document representations, such as the use of word order and syntactical and/or semantical information. We intend to explore this as future work besides trying other feature selection techniques.

Another characteristic of the datasets that may affect the SVM performance is the class imbalance of the binary classifiers – there are much more negative examples than positive ones. As referred in [3] this can be a source of bad results. We intend, also, to address this problem as future work.

References

1. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
2. J. Furnkranz, T. Mitchell, and E. Riloff. A case study in using linguistic phrases for text categorization on the www. In *Proceedings of AIII/ICML-98 Workshop on Text Categorization*, pages 5–12, 1999.
3. N. Japkowicz. The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000)*, volume 1, pages 111–117, 2000.
4. T. Joachims. *Learning to Classify Text Using Support Vector Machines*. Kluwer academic Publishers, 2002.
5. D. Mladenic and M. Grobelnik. Feature selection for unbalanced class distribution and naive bayes, 1999.
6. M. Porter. An algorithm for suffix stripping. *Program (Automated Library and Information Systems)*, 14(3):130–137, 1980.
7. P. Quaresma and I. 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.
8. G. Salton and M. McGill. *Introduction to Modern Informatin Retrieval*. McGraw-Hill, 1983.
9. B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.
10. S. Scott and S. Matwin. Feature engineering for text classification. In *Proceedings of ICML-99, 16th International Conference on Machine Learning*, pages 379–388, 1999.
11. V. Vapnik. *The nature of statistical learning theory*. Springer, 1995.
12. I. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, 1999.