Using IR techniques to improve Automated Text Classification

Teresa Gonçalves and Paulo Quaresma

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

Abstract. This paper performs a study on the pre-processing phase of the automated text classification problem. We use the linear Support Vector Machine paradigm applied to datasets written in the English and the European Portuguese languages – the Reuters and the Portuguese Attorney General's Office datasets, respectively.
The study can be seen as a search, for the best document representation, in three different axes: the feature reduction (using linguistic information), the feature selection (using word frequencies) and the term weighting (using information retrieval measures).

1 Introduction

In the last years text classification is gaining popularity due to the increased availability of documents in digital form and the following need to access them in flexible ways. This problem is well known in the Information Retrieval community and the use of Machine Learning techniques is opening many important and interesting research problems.

Research aimed at the application of Machine Learning methods to text classification has been conducted among others by Apté et al. (rule-based induction methods) [1], Mladenić and Grobelnik (naïve Bayes) [7], Nigam et al. (EM and naïve Bayes) [6] and Joachims (SVM – support vector machines) [5].

In Joachims's work, documents are represented as bag-of-words [9] (without word order information) and the results are evaluated using information retrieval measures, such as the *precision recall break-even point* (PRBP).

In this paper, we follow his approach, aiming to determine if linguistic information is helpful for achieving good SVM performance. We use two sets of documents written in two different languages – the European Portuguese (the PAGOD dataset [8]) and the English one (the Reuters dataset).

The work can be seen as a search in three different axes: the feature reduction (using linguistic information), the feature selection (using word frequencies) and the term weighting (using information retrieval measures) axes.

On previous work, we evaluated SVM performance compared with other Machine Learning algorithms [2] and performed a preliminary study on the impact of using linguistic information to reduce the number of features [3]. In this paper, we extend that work using IR techniques to weight and normalise features.

In Section 2 a brief description of the Support Vector Machines theory is presented, while in Section 3 our classification problem and datasets are characterised. Our experiments are described in Section 4 and the results are presented in Section 5. Finally, some conclusions and future work are pointed out in Section 6.

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 binary linear 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, the functions that transform the input feature space, are used. More detailed information can be obtained in several specialised books, such as [10].

3 Domain Description

The text classification problem at hand (both, the Reuters and the PAGOD datasets), can be characterised as a multi-label one, i.e. documents can be classified into multiple concepts/topics. The typical approach to solve it, is to divide into a set of binary problems, where each concept is considered independently, reducing the initial problem to several binary classification ones.

An important open problem is the representation of the documents. In this work, as already mentioned, we will use the standard vector representation, where each document is represented as a bag-of-words. We discarded all words containing digits and retained words' frequencies.

3.1 The 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 and 3299 testing documents.

On all 12902 documents, we found 31715 distinct words; per document, we obtained averages of 126 words, of which 70 were distinct.

3.2 The PAGOD dataset

This dataset has 8151 documents and represent the decisions of the Portuguese Attorney General's Office since 1940. It is written in the European Portuguese language, and delivers 96 MBytes 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.

From all potential categories, a preliminary evaluation showed that only about 3000 terms were. We found 68886 distinct words and, per document, we obtained averages of 1339 words, of which 306 were distinct.

4 Experiments

We chose the top five concepts and applied the SVM learning algorithm using a linear kernel. For each dataset we performed three classes of experiments: a feature reduction one (using linguistic information), a rudimentary kind of feature selection and some term weighting techniques (from the IR field). For each experiment we analysed the precision, recall and F_1 measures [9].

We generated a linear SVM for each possible combination of the experiments' classes, using the WEKA package [12] from Waikato University, with default parameters. For the Reuters dataset we used the training and test sets, while for the PAGOD dataset we performed a 10-fold cross validation procedure.

4.1 Feature Reduction

On trying to reduce the number of features we made three different experiments: in rdt_1 we used no linguistic information, in rdt_2 we removed a list of considered non-relevant words (such as articles, pronouns, adverbs and prepositions) and in rdt_3 we removed the same non-relevant words and transformed each remaining word onto its stem (its lemma for the Portuguese language).

In the Reuters dataset we used the FreeWAIS stop-list to remove the non-relevant words and the Porter algorithm to transform each word onto its stem. In the PAGOD dataset, this work was done using a Portuguese lexical database, POLARIS, that provided the lemmatisation of every Portuguese word.

4.2 Feature Selection

Feature selection was done by eliminating the words that appear less than a specific number in the set of all documents: for example, sel_{55} means that all words that appeared less than 55 times in all documents were eliminated. We performed experiences for sel_1, sel_{50}, sel_{100}, sel_{200}, sel_{400}, sel_{800} and sel_{1600}.

4.3 Term Weighting

Term weighting techniques usually consist of three components: the document, the collection and the normalisation components [9]. For the final feature vector x, the value x_i for word w_i is computed by multiplying the three components.

We tried four different combinations of components: wgt_1 is the *binary representation* with no collection component but normalised to unit length; wgt_2 uses the raw term frequencies (TF) with no collection component nor normalisation; wgt_3 uses TF with no collection component but normalised to unit length; wgt_4 is the popular $TFIDF$ representation (TF divided by the document frequency, DF, i.e. the number of documents in which w_i occurs at least once) normalised to unit length.

5 Results

For reasons of space we only show,for each dataset, the values obtained for the micro and macro averaging of the F_1-*measure*.

		micro				**macro**			
		wgt_1	wgt_2	wgt_3	wgt_4	wgt_1	wgt_2	wgt_3	wgt_4
red_1	sel_1	0.926	0.891	0.930	0.932	0.859	0.792	0.874	0.874
	sel_{50}	0.918	0.905	0.933	0.930	0.855	0.823	0.887	0.886
	sel_{100}	0.919	0.898	0.933	0.928	0.858	0.798	0.882	0.880
	sel_{200}	0.920	0.894	0.929	0.930	0.858	0.792	0.873	0.876
	sel_{400}	0.920	0.888	0.923	0.924	0.858	0.767	0.855	0.859
	sel_{800}	0.897	0.860	0.898	0.901	0.808	0.700	0.800	0.809
	sel_{1600}	0.854	0.809	0.855	0.849	0.726	0.558	0.703	0.690
red_2	sel_1	0.926	0.888	0.931	0.928	0.863	0.790	0.876	0.869
	sel_{50}	0.920	0.905	0.936	0.929	0.866	0.823	0.888	0.882
	sel_{100}	0.923	0.899	0.937	0.935	0.872	0.806	0.886	0.889
	sel_{200}	0.923	0.897	0.936	0.932	0.866	0.800	0.885	0.877
	sel_{400}	0.924	0.884	0.927	0.926	0.867	0.760	0.865	0.862
	sel_{800}	0.895	0.844	0.893	0.889	0.813	0.680	0.799	0.795
	sel_{1600}	0.841	0.759	0.833	0.832	0.721	0.488	0.622	0.624
red_3	sel_1	0.923	0.889	0.936	0.930	0.861	0.799	0.882	0.874
	sel_{50}	0.920	0.902	0.934	0.931	0.862	0.824	0.882	0.878
	sel_{100}	0.924	0.900	0.937	0.937	0.868	0.813	0.889	0.891
	sel_{200}	0.921	0.898	0.935	0.933	0.866	0.806	0.884	0.878
	sel_{400}	0.921	0.886	0.932	0.928	0.862	0.766	0.873	0.864
	sel_{800}	0.914	0.863	0.913	0.910	0.839	0.708	0.821	0.815
	sel_{1600}	0.844	0.786	0.852	0.845	0.698	0.521	0.689	0.641

Table 1. Micro and macro F_1 for the Reuters dataset.

Analysing Reuters' results (Table 1), one can say that, for the feature selection axis, the wgt_3 experiment presents the best F_1 values (both maximum and average values). On the feature reduction axis, and taking into account the previous choice, red_3 is the best experiment. Finally, and for the remaining axis, the sel_{100} experiment is the one that presents the best values. These choices are valid for both macro and micro averaging.

On the other hand, for the PAGOD dataset (Table 2) and using the same procedure, wgt_1 and wgt_3 present best results for the feature selection axis and

		micro				macro			
		wgt_1	wgt_2	wgt_3	wgt_4	wgt_1	wgt_2	wgt_3	wgt_4
red_1	sel_1	0.759	0.687	0.732	0.722	0.652	0.531	0.631	0.620
	sel_{50}	0.750	0.694	0.694	0.678	0.651	0.509	0.601	0.587
	sel_{100}	0.747	0.692	0.712	0.700	0.652	0.497	0.615	0.604
	sel_{200}	0.744	0.694	0.731	0.720	0.649	0.502	0.634	0.619
	sel_{400}	0.734	0.688	0.743	0.737	0.644	0.485	0.641	0.629
	sel_{800}	0.730	0.659	0.746	0.740	0.635	0.464	0.638	0.620
	sel_{1600}	0.745	0.579	0.754	0.744	0.642	0.402	0.632	0.606
red_2	sel_1	0.760	0.687	0.757	0.754	0.660	0.533	0.659	0.654
	sel_{50}	0.750	0.697	0.735	0.737	0.652	0.514	0.640	0.640
	sel_{100}	0.750	0.692	0.740	0.738	0.655	0.500	0.646	0.643
	sel_{200}	0.745	0.691	0.747	0.748	0.656	0.497	0.655	0.655
	sel_{400}	0.740	0.690	0.756	0.756	0.650	0.488	0.661	0.656
	sel_{800}	0.743	0.659	0.754	0.750	0.644	0.467	0.650	0.633
	sel_{1600}	0.754	0.574	0.763	0.749	0.645	0.399	0.646	0.610
red_3	sel_1	0.751	0.673	0.752	0.747	0.652	0.493	0.658	0.653
	sel_{50}	0.751	0.677	0.746	0.740	0.656	0.480	0.654	0.647
	sel_{100}	0.744	0.672	0.748	0.740	0.650	0.474	0.655	0.643
	sel_{200}	0.742	0.674	0.754	0.749	0.649	0.475	0.661	0.651
	sel_{400}	0.740	0.671	0.761	0.758	0.647	0.473	0.667	0.656
	sel_{800}	0.750	0.631	0.759	0.756	0.652	0.449	0.655	0.637
	sel_{1600}	0.743	0.560	0.758	0.745	0.630	0.398	0.638	0.603

Table 2. Micro and macro F_1 for the PAGOD dataset.

red_2 and red_3 are the best experiments for the feature reduction one. Concerning the feature selection, it is not possible to get a winning experiment. These results are also valid for the micro and macro averaging F_1 values.

6 Conclusions and Future Work

From the previous section and for both datasets, one can reason out that the best term weighting technique is the one that counts term frequencies and normalises it to unit length. From feature reduction results, one can say that linguistic information is useful for getting better performance.

Concerning the feature selection experiments, it is not possible to reach a conclusion valid for both datasets: for the Reuters we have a winning experiment (sel_{100}) while for the PAGOD we have not. This can a characteristic of the written language or of the documents by themselves (for example, on average, the Reuters documents are shorter that the PAGOD ones). Nevertheless, it is possible to say that one can build better classifiers, quicker without loosing performance. Just as an example, and for the PAGOD dataset we are talking on a reduction from almost 6 hours (for sel_1) to 1 hour and half (sel_{400}) for the wgt_3–rdt_1 experiment.

As future work, we intend to add another axis on our study: the selection of the best features that describe each concept. Instead of word frequencies, we intend to use other measures, like the Mutual Information from Information Theory.

We also intend to study the impact of the imbalance nature of these datasets on the SVM performance. In fact, there are much more negative examples than positive ones on the binary classifiers and this can be a source of bad results as referred for instance in [4].

Going further on our future work, we intend to address the document representation problem, by trying more powerful representations than the bag-of-words used in this work. Aiming to develop better classifiers, we intend to explore the use of word order and the syntactical and/or semantical information on the representation of documents.

References

1. C. Apté, F. Damerau, and S. Weiss. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3):233–251, 1994.
2. T. Gonçalves and P. Quaresma. A preliminary approach to the multi-label classification problem of Portuguese juridical documents. In *11th Portuguese Conference on Artificial Intelligence*, Lecture Notes on Artificial Intelligence 2902, pages 435–444, Évora, Portugal, December 2003. Springer-Verlag.
3. Teresa Gonçalves and Paulo Quaresma. The impact of NLP techniques in the multilabel classification problem. In *Intelligent Information Systems 2004*, Advances in Soft Computing, Zakopane, Poland, May 2004. Springer-Verlag. (to appear).
4. 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.
5. T. Joachims. *Learning to Classify Text Using Support Vector Machines.* Kluwer Academic Publishers, 2002.
6. K.Nigam, A.McCallum, S.Thrun, and T.Mitchell. Text classification from labelled and unlabelled documents using EM. *Machine Learning*, 39(2):103–134, 2000.
7. D. Mladenić and M. Grobelnik. Feature selection for unbalanced class distribution and Naïve Bayes. In *Proceedings of ICML-99, 16th International Conference on Machine Learning*, pages 258–267, 1999.
8. 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.
9. G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.
10. B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, 2002.
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.