Using Micro-Documents for Feature Selection: The Case of Ordinal Text Classification

Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani

Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, Italy
E-mail: {firstname.lastname}@isti.cnr.it

Abstract. Most popular feature selection methods for text classification (TC) are based on binary information concerning the presence/absence of the feature in each training document. As such, these methods do not exploit term frequency information. In order to overcome this drawback we break down each training document of length k into k training "micro-documents", each consisting of a single word occurrence and endowed with the class information of the original training document. We study the impact of this strategy in the case of ordinal TC; the experiments show that this strategy substantially improves effectiveness.

1 Introduction

Feature selection (FS) is a technique for reducing the dimensionality of a vector space in learning tasks (see e.g., [1]). It consists in identifying a subset $S \subset T$ of the original feature set T such that $|S| \ll |T|$ (with $\xi = |S|/|T|$ called the *reduction level*) and such that S reaches the best compromise between (a) the efficiency of the learning process and of the classifiers, and (b) the effectiveness of the resulting classifiers. In text classification (TC) the most popular approach to FS is the *filter* approach: a real-valued function f is applied to each feature in T in order to compute its expected contribution to solving the classification task, and only the $|S|$ features with the highest f value are retained.

The most popular instances of function f, such as information gain (a.k.a. mutual information), chi-square, odds ratio, pointwise mutual information, and the like, are based on *binary* information concerning the *presence/absence* of the feature in each training document. For instance, in pointwise mutual information, defined as $PMI(t_k, c_j) = \log_2 \frac{P(t_k,c_j)}{P(t_k)P(c_j)}$, the value $P(t_k)$ is the probability that feature t_k occurs at all in a random training document. As such, PMI and all the other above-mentioned functions do not exploit a rich source of information, namely, the information concerning *how many times* t_k occurs in a given training document (*term frequency*). In this paper we propose a filter approach to FS which attempts to overcome this drawback. The approach consists in breaking down each training document d_i into $length(d_i)$ training "micro-documents" (μ-*documents*), each consisting of a single word occurrence and endowed with the same class information of the original training document. In this paper we limit our experiments to the case of "ordinal" text classification (see below).

This paper is organized as follows. In Section 2 we present our μ-document-based approach to FS. Section 3 describes experiments we have conducted using two SVM-based learning methods and two large datasets of product reviews.

2 Feature Selection for OC based on Training μ-documents

Let us fix some terminology and notation. *Ordinal classification* (OC – also known as *ordinal regression*) consists in estimating (from a training set Tr) a *target function* $\Phi : X \rightarrow R$ which maps each object $x_i \in X$ into exactly one of an ordered sequence (here called *rankset*) $R = \langle r_1 \prec \ldots \prec r_n \rangle$ of *ranks* (aka "scores", or "labels", or "classes"). The result of the estimation is a function $\hat{\Phi}$ called the *classifier*, which we will evaluate on a test set Te. Our FS methods will typically consist of (a) scoring each feature $t_k \in T$ by means of a function that measures the predicted utility of t_k for the classification process, and, (b) given a predetermined reduction level ξ, selecting the $|S| = \xi \cdot |T|$ top-scoring features.

The FS methods that we use in this paper are the $Var{*}IDF$, $RR(Var{*}IDF)$, $RR(IGOR)$ and $RR(AC * IDF)$ methods originally defined in [2] (an extended version of [3]), to which we refer the reader for details. All these functions only use information concerning the presence/absence of feature t_k in training document d_i. We attempt to overcome this drawback by breaking down each training document d_i into $length(d_i)$ training "μ-documents", each consisting of a single word occurrence and endowed with the same class information of the original training document. The training set Tr is then replaced, *for FS purposes only*, by the set of the training "μ-documents" obtained from it. All the original FS methods are obviously still applicable after this move: however, these methods are now *de facto* sensitive to term frequency, since a training document d_j belonging to class c_i and containing r occurrences of feature t_k has generated (among others) r training μ-documents containing (only) t_k and belonging to c_i.

The move from training documents to training μ-documents is, as far as FS is concerned, akin to the move, in naïve Bayesian learners, from a multivariate Bernoulli event model (where documents are events) to a multinomial event model (where word occurrences are events). In the context of TC this move was originally discussed in [4]. However, in that case the authors reported that little difference in performance was found when selecting features via the former model rather than via the latter model (no actual effectiveness figures were given, though). Our work may be seen as exporting that idea outside the realm of naïve Bayesian learners, and outside the realm of single-label TC, neither of which has been done before to the best of our knowledge.

3 Experiments

We have tested the proposed method on two different datasets for ordinal text classification. The first is the TripAdvisor-15763 dataset, with the same split between training and test documents as used in [2], resulting in 10,508 documents used for training and 5,255 for test. The second dataset is the Amazon-83713, with the same split between training and test documents as in [2], resulting in 20,000 documents used for training and 63,713 for test. Both datasets consist of textual product reviews scored on a scale from 1 to 5 "stars". As our main evaluation measure we use the *macroaveraged mean absolute error* (MAE^M) measure proposed in [5].

We have tested our methods with two different SVM-based learning algorithms for ordinal regression, ϵ-SVR and SVOR; see [2] for more details. As the baselines against which to test our μ-documents-based approach we have used the results we have obtained in [2] (on the same datasets and with the same learning algorithms) with the versions based on "regular" training documents of the $Var*IDF$, $RR(Var*IDF)$, $RR(IGOR)$ and $RR(AC*IDF)$ methods. We have set the γ and C parameters of both learners to the optimal values that we had obtained in the experiments of [2]; this means that the parameters are optimal for the baselines but not necessarily for the methods proposed here, which lends even higher value to the results obtained by the methods proposed here.

The experimental protocol essentially conforms to that of [2]. As a vectorial representation, after stop word removal (and no stemming) we have used standard bag-of words with cosine-normalized $tfidf$ weighting. We have run all our experiments for all the 100 reduction levels $\xi \in \{0.001, 0.01, 0.02, 0.03, \ldots, 0.99\}$.

For the $Var*IDF$, $RR(Var*IDF)$ and $RR(AC*IDF)$ methods we have set the smoothing parameter ϵ (see [2]) to 0.1. For the same methods we have used the optimal (individually for each method) values of the a parameter (see [2]) that we had obtained in the experiments of [2]; again, this means that the parameters are optimal for the baselines but not necessarily for the methods proposed here. For $RR(AC*IDF)$, the E error measure (see [2]) was taken to be $|\hat{\Phi}(d_i) - \Phi(d_i)|$ (i.e., absolute error), given that it is the document-level analogue of MAE^M.

3.1 Results

The main observation to be made from the results of our experiments (which are not reported here in detail reasons of space – see [6] for more details) is that the use of training μ-documents substantially enhances the accuracy of ordinal TC, since it is practically always the case that the MAE^M values of the μ-document-based versions are better than the corresponding values of the "regular document" -based versions, irrespective of FS function, dataset, and learner. Overall, these results derive from a massive experimental work, consisting of (100 reduction levels × 2 datasets × 2 learners × 4 FS functions =) 1600 new train-and-test experiments (which are additional to the other 1600 that produced out baselines and that were already presented in [2]).

	TripAdvisor-15763						Amazon-83713					
	ϵ-SVR			SVOR			ϵ-SVR			SVOR		
	RDs	μDs	Δ	RDs	μDs	Δ	RDs	μDs	Δ	RDs	μDs	Δ
$Var * IDF$	0.722	0.658	(-8.84%)	0.818	0.787	(-3.82%)	0.691	0.645	(-6.69%)	0.818	0.790	(-3.51%)
$RR(Var * IDF)$	0.688	0.660	(-4.10%)	0.797	0.787	(-1.36%)	0.694	0.644	(-7.26%)	0.833	0.821	(-1.52%)
$RR(IGOR)$	0.695	0.665	(-4.25%)	0.800	0.800	(-0.00%)	0.689	0.646	(-6.23%)	0.838	0.837	(-0.12%)
$RR(AC * IDF)$	0.680	0.666	(-2.09%)	0.818	0.780	(-4.71%)	0.697	0.662	(-4.99%)	0.837	0.787	(-5.92%)
Average	0.696	0.662	(-4.87%)	0.808	0.788	(-2.48%)	0.693	0.649	(-6.29%)	0.831	0.808	(-2.77%)

Table 1. Average values of MAE^M computed across the 100 different values for ξ; Δ indicates the relative reduction in average MAE^M obtained by replacing regular training documents (RDs) with μ-documents (μDs).

Table 1 reports MAE^M values as averaged, for a given combination of dataset and learner, across the 100 values of ξ. A further interesting observation that this table allows to draw is that the improvements brought about by the μ-documents technique are much higher for ϵ-SVR than for SVOR; the fact that ϵ-SVR is practically always a better performer than SVOR lends thus to the μ-documents technique even higher value.

4 Conclusion

We have shown that using μ-documents in place of "regular" training documents in feature selection substantially improves the effectiveness of ordinal text classification. In future experiments we plan to validate this method on the more standard cases of binary classification and multiclass classification.

References

1. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference on Machine Learning (ICML'97), Nashville, US (1997) 412–420
2. Baccianella, S., Esuli, A., Sebastiani, F.: Feature selection for ordinal text classification. Technical Report 2010-TR-014, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT (2010)
3. Baccianella, S., Esuli, A., Sebastiani, F.: Feature selection for ordinal regression. In: Proceedings of the 25th ACM Symposium on Applied Computing (SAC'10), Sierre, CH (2010) 1748–1754
4. McCallum, A.K., Nigam, K.: A comparison of event models for naive Bayes text classification. In: Proceedings of the AAAI Workshop on Learning for Text Categorization, Madison, US (1998) 41–48
5. Baccianella, S., Esuli, A., Sebastiani, F.: Evaluation measures for ordinal text classification. In: Proceedings of the 9th IEEE International Conference on Intelligent Systems Design and Applications (ISDA'09), Pisa, IT (2009) 283–287
6. Baccianella, S., Esuli, A., Sebastiani, F.: Using micro-documents for feature selection: The case of ordinal text classification. Technical Report 2011-TR-001, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT (2011)