Information Retrieval 1, 69–90 (1999)

An Evaluation of Statistical Approaches to Text Categorization*

YIMING YANG yiming@cs.cmu.edu
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213-3702, USA

Received October 28, 1997; Revised May 13, 1998; Accepted July 27, 1998

Abstract. This paper focuses on a comparative evaluation of a wide-range of text categorization methods, including previously published results on the Reuters corpus and new results of additional experiments. A controlled study using three classifiers, kNN, LLSF and WORD, was conducted to examine the impact of configuration variations in five versions of Reuters on the observed performance of classifiers. Analysis and empirical evidence suggest that the evaluation results on some versions of Reuters were significantly affected by the inclusion of a large portion of unlabelled documents, mading those results difficult to interpret and leading to considerable confusions in the literature. Using the results evaluated on the other versions of Reuters which exclude the unlabelled documents, the performance of twelve methods are compared directly or indirectly. For indirect compararions, kNN, LLSF and WORD were used as baselines, since they were evaluated on all versions of Reuters that exclude the unlabelled documents. As a global observation, kNN, LLSF and a neural network method had the best performance; except for a Naive Bayes approach, the other learning algorithms also performed relatively well.

Keywords: text categorization, statistical learning algorithms, comparative study, evaluation

1. Introduction

Text categorization (TC) is the problem of assigning predefined categories to free text documents. A growing number of statistical learning methods have been applied to this problem in recent years, including regression models (Fuhr et al. 1991, Yang and Chute 1994), nearest neighbor classifiers (Creecy et al. 1992, Yang 1994), Bayesian probabilistic classifiers (Tzeras and Hartman 1993, Lewis and Ringuette 1994, Moulinier 1997), decision trees (Fuhr et al. 1991, Lewis and Ringuette 1994, Moulinier 1997), inductive rule learning algorithms (Apte et al. 1994, Cohen and Singer 1996, Moulinier et al. 1996), neural networks (Wiener et al. 1995, Ng et al. 1997) and on-line learning approaches (Cohen and Singer 1996, Lewis et al. 1996). With more and more methods available, cross-method evaluation becomes increasingly important to identify the state-of-the-art in text categorization. However, without a unified methodology in empirical evaluations, objective comparisons of different methods are difficult.

Ideally, all researchers would like to use a common collection and comparable performance measures to evaluate their systems, or would allow their systems to be evaluated under carefully-controlled conditions in a fashion similar to that of the Text Retrieval Conference

*This research was supported in part by NIH grant LM-05714 and by NSF grant IRI9314992.

(TREC). The reality, however, is far from the ideal. Cross-method comparisons have been attempted in the literature, but often only for two or three methods. The small scale of these comparisons could either lead to overly-general statements based on insufficient observations, or provide little insight into a global comparison between a wide range of approaches. An alternative to these small-scale comparisons would be to integrate the available results of categorization methods into a global evaluation, carefully analyzing the test conditions and evaluation measures used and establishing a common basis for cross-collection and cross-experiment integration. This solution would lead to a TREC-like controlled evaluation for text categorization, as well as contribute useful insights to individual studies. This paper is an effort in that direction.

The most serious problem in TC evaluations is the lack of standard data collections. Even if a common collection is chosen, there are still many ways to introduce inconsistent variations. The commonly used Reuters news story corpus, for example, has at least five different versions, depending on how the training/test sets are divided, which subsets of categories or documents were used or not used for evaluation, and so forth. The number of different configurations of this corpus is still growing. It is often unclear whether or not the reported results on the different versions of Reuters are comparable to each other. In this paper we examine the impact of corpus configuration variations on the performance of classifiers, using a carefully-controlled experiments of several categorization systems on five different versions of Reuters. As will be shown in Section 5.2, variations between certain versions of Reuters *do* have a strong impact, while the variations between other versions do not. The underlying reason for this will be analyzed.

Another important issue in cross-experiment evaluation is the comparability between different performance measures used in individual experiments. Many measures have been used, including recall and precision, accuracy or error, *break-even point* or *F-measure*, *micro-average* and *macro-average* for binary categorization, *11-point average precision* for category ranking, and so forth (see Section 3 for definitions). Each of these measures is designed to evaluate some aspect of the categorization performance of a system; however, none of them convey identical information. Which of these measures are more suitable for text categorization? How can published results of text categorization methods be best compared when they were evaluated using different performance measures? These questions are addressed in this paper by applying a variety of performance measures to several classifiers, including the measures for category ranking evaluation or the measures for binary category assignment. We will show that both types of measures are informative and complementary to each other. We will also show that with carefully chosen performance measures and a baseline classifier, one can reasonably (indirectly) compare the relevant performance among classifiers across experiments based on their relevant performance with respect to the baseline classifier.

This paper is divided into six sections in addition to the introduction. Section 2 describes the classifiers and the Reuters corpus we will use in this paper. Section 3 introduces and analyzes performance measures for category ranking evaluation and binary categorization evaluation. Section 4 describes the novel experiments we conducted with WORD, kNN, and LLSF. Section 5 reports the results of our classifiers and evaluate them together with published results of other classifiers. Finally, we summarize our conclusions in Section 6.