Comparison of Word-based and Letter-based Text Classification

Victoria Bobicev
Technical University of Moldova
Studentilor, 7, Chisinau, Moldova
victoria_bobicev@rol.md

Abstract

In this paper the comparison of two PPM (Prediction by Partial Matching) methods for automatic content-based text classification is described: on the basis of letters and on the basis of words.

The investigation was driven by the idea that words and especially word combinations are more relevant features for many text classification tasks than letters and letter combinations. The results of the experiments proved applicability of PPM models for content-based text classification, although PPM model on the basis of words did not perform better than model on the basis of letters.

1. Introduction

Text or document classification is the assignment of documents to predefined categories on the base of their content.

In this paper the application of word-based PPM (Prediction by Partial Matching) model for automatic content-based text classification is explored. Although the application of PPM model to the document classification is not new, all the PPM models used for text classification were character-based and used sequences of two or more letters as features [20]. On the other hand, typical approaches to text classification use words as features for feature vector creation.

The main idea investigated in the paper is that words and especially word combinations are more relevant features for many text classification tasks. It is known that key-words for a document in most cases are not just a single word but combination of two or three words. That is why word-based PPM model was created and used for text classification.

2. Related Works

A wide variety of learning approaches to text categorisation have been used, including Bayesian classification [6], decision trees [15], cluster classification [12], k-NN algorithms [5] and neural nets [17]. Lately the most wide spread classification techniques are based on the SVM (support vector machine) [11].

Several approaches that apply compression models to text classification have been presented recently [2], [7], [21]. The underlying idea of using compression methods for text classification was their ability to create the language model adapted to particular texts. It was supposed that this model captures individual features of the text being modelled. Theoretical background to this approach was given in [20].

3. PPM Compression

PPM (prediction by partial matching) is an adaptive finite-context method for compression. It is based on probabilities of the upcoming symbol in dependence of several previous symbols. Firstly this algorithm was presented in [3], [4]. Lately the algorithm was modified and an optimized PPMC (Prediction by Partial Matching, escape method C) algorithm was described in [16]. PPM has set the performance standard for lossless compression of text throughout the past decade. The PPM technique blends character context models of varying length to arrive at a final overall probability distribution for predicting upcoming characters in the text.

For example, the probability of character '***m***' in context of the word '***algorithm***' is calculated as a sum of conditional probabilities in dependence of different length context up to the limited maximal length:

$$P_{PPM}(\text{'}\boldsymbol{m}\text{'}) = \lambda_5 \cdot P(\text{'}\boldsymbol{m}\text{'} \mid \text{'}\boldsymbol{orith}\text{'}) + \lambda_4 \cdot P(\text{'}\boldsymbol{m}\text{'} \mid \text{'}\boldsymbol{rith}\text{'}) +$$
$$+ \lambda_3 \cdot P(\text{'}\boldsymbol{m}\text{'} \mid \text{'}\boldsymbol{ith}\text{'}) + \lambda_2 \cdot P(\text{'}\boldsymbol{m}\text{'} \mid \text{'}\boldsymbol{th}\text{'}) + \lambda_1 \cdot P(\text{'}\boldsymbol{m}\text{'} \mid \text{'}\boldsymbol{h}\text{'}) +$$
$$+ \lambda_0 \cdot P(\text{'}\boldsymbol{m}\text{'}) + \lambda_{-1} \cdot P(\text{'esc'}),$$

where λ_i (i = 1…5) is normalization factor;
5 - maximal length of the context;

P('esc') – 'escape' probability, the probability of the character that have never been encountered so far.

4. Classification Using PPM Models

Most of compression models are character-based. They treat the text as a string of characters. This method has several potential advantages. For example, it avoids the problem of defining word boundaries; it deals with different types of documents in a uniform way. It can work with text in any language and it can be applied to diverse types of classification.

In [14] the simplest way of compression-based categorization called 'off-the-shelf algorithm' is used for authorship attribution. The main idea of this method is as follows. Anonymous text is attached to texts which characterize classes, and then it is compressed. A model, providing the best compression of document, is considered as having the same class with it.

The other approach is direct measuring of text entropy using a certain text model. PPM is appropriate in this case, because text modelling and its statistic encoding are two different stages in this method. In [13] was shown that results of this method were very similar to the results of the 'off-the-shelf algorithm'.