Abstract

Current approaches to feature selection for text classification aim to reduce the number of terms that are used to describe documents. Thus, documents can be classified and found with
greater ease and precision. A key shortcoming of these approaches is that they select the topmost terms to describe documents after ranking all terms using a feature selection measure (scoring function). Lesser high-ranking terms below the topmost terms are discarded to reduce computational costs. Nevertheless, in many cases, they may have considerable discriminative power to enhance the text classification precision. In order to address this issue, we proposed a new feature weighting formalism that ties the topmost terms with lesser high-ranking terms using probabilistic neural networks. In the proposed method, probabilistic neural networks are formed using relative category distribution matrix and topmost terms are re-weighted and passed to Rocchio classifier. This is achieved without increasing the dimensionality of the feature space. Through experiments on datasets from Reuters news collection RCV1, we show that the proposed method is a significant supplement to the statistical feature selection measures for better text classification at extreme term filtering ranges.

Reference


Index Terms

Computer Science  Pattern Recognition

Key words

Term re-weighting  boosting  probabilistic
neural networks

text classification

feature selection

Rocchio classifier