KYBERNETIKA — VOLUME *40* (2004), NUMBER 3, PAGES 293-304

Text Document Classification Based on Mixture Models

Jana Novovičová; Antonín Malík

Abstract: Finite mixture modelling of class-conditional distributions is a standard method in a statistical pattern recognition. This paper, using bag-of-words vector document representation, explores the use of the mixture of multinomial distributions as a model for class-conditional distribution for multiclass text document classification task. Experimental comparison of the proposed model and the standard Bernoulli and multinomial models as well as the model based on mixture of multivariate Bernoulli distributions was performed using Reuters-21578 and Newsgroups data sets. Preliminary experimental results indicate the effectiveness of the proposed model in a text classification problem.

Keywords: text classification; multinomial mixture model;

AMS Subject Classification: 62H30; 62G05; 68T10;