Intelligent Automation and Soft Computing, Vol. 16, No. 1, pp. 25-38, 2010

DOCUMENT CLASSIFICATION USING INFORMATION THEORY AND A FAST BACK-PROPAGATION NEURAL NETWORK

Howard Li[1]*, Liam Paull[1], Yevgen Biletskiy[1], and Simon X. Yang[2]

[1]Department of Electrical and Computer Engineering
University of New Brunswick
Fredericton, New Brunswick E3B 5A3, Canada

[2]School of Engineering, University of Guelph
Guelph, Ontario N1G 2W1, Canada

ABSTRACT—In this paper, a fast back-propagation neural network is developed to build document classifiers and the information gain method is used for feature selection. According to the rank of the information gain of all the words contained in the documents, those words that contain more information to classify the documents are selected as the input features of the artificial neural network (ANN) classifiers. The neural network developed assumes a three-layer structure with a fast back-propagation learning algorithm. Because of the information contained in the vectors selected, the learning efficiency of the developed ANN is very high. For the output of the ANN, Shannon entropy is used to tune the threshold of the binary classifiers. The classifiers are tested using the Reuters corpus. Two performance measures are used to evaluate the performance of the classifiers and generally the results of this study are better than those claimed in literature.