Jordan University of Science and Technology

An Improved Classifier for Arabic Text

Authors: Amer Al-Badarneh, Mostafa Ali, Safwan Ghaleb

Abstract: K-Nearest Neighbor text classifier is considered a lazy technique since it compares the target document with all documents in the training set and the classifier treats all training documents the same, even so some documents are more important than others. This paper introduces an enhanced k-NN algorithm for Arabic text categorization based on clustering. The proposed approach makes use of KMeans clustering algorithm to cluster the dataset and then take the new centers of the clustered data to be the new training set. In order to eliminate the multi-peak effect and improve the accuracy of the proposed approach, the documents that are near a calculated border are removed. To overcome the biased, resulting from the clustering process, we use a weighting scheme that assigns a weight value for each cluster center based on the number of documents in that cluster. Finally, we apply the k-NN classifier on the clustered and weighted data. Our technique was tested on a large benchmark dataset that includes documents from different categories. The results show that our technique has outperformed several other well-known classifiers.