CLASS FEATURE CENTROID CLASSIFIER METHOD FOR
CLASSIFYING IMBALANCED ARABIC TEXT

Name : Indriati
Registration Number : 5110 201 010
Advisory : Dr. Agus Zainal Arifin, S.Kom, M.Kom
Anny Yuniarti, S.Kom, M.Comp.Sc

ABSTRACT

The large amount of unstructured digital texts such as an Arabic text required an automatic text classification. Many previous researches in this field incorporated dataset that had been created themselves therefore the quality of the result is imbalance and difficult to be determined. This research proposes a method for text classification using the Class Feature Centroid (CFC) by adding Imbalanced Factor (IF) on weighted term. This method is used for imbalanced data in a binary class with the Arabic text document. The Arabic text documents are processed in the preprocessing stage to generate the terms. In the training phase, these terms will be weighted by the CFC method by adding IF, which will form the centroid of each class. In the testing phase, the centroid is used to classify the documents. The document with the greatest similarity to a centroid in a class will be a member of that class. From the experiments, the best result of F Measure is 0,97, while using the original CFC method, the best result of F Measure is 0,95. The result of experiments indicates that the CFC method by adding IF can be used for classifying imbalanced Arabic text data.

Keywords: text classification, centroid-based classification, Class Feature Centroid, Imbalanced Factor