Jordan University of Science and Technology

Clustering Generalised Instances Set Approaches

Authors: Hassan Najadat, Rasha Obeidaty, and Ismail Hmeidi

Abstract: This paper introduces three new text classi?cation methods: Clustering-Based Generalised Instances Set (CB-GIS), Multilevel Clustering-Based Generalised Instances Set (MLC GIS) and Multilevel Clustering-Based k Nearest Neighbours (MLC-kNN). These new methods aim to unify the strengths and overcome the drawbacks of the three similaritybased text classi?cation methods, namely, kNN, centroid-based and GIS. The new methods utilise a clustering technique called spherical K-means to represent each class by a representative set of generalised instances to be used later in the classi?cation. The CB-GIS method applies a ?at clustering method while MLC-GIS and MLC-kNN apply multilevel clustering. Extensive experiments have been conducted to evaluate the new methods and compare them with kNN, centroid-based and GIS classi?ers on the Reuters-21578(10) benchmark dataset. The evaluation has been performed in terms of the classi?cation performance and the classi?cation e?ciency. The experimental results show that the top-performing classi?cation method is the MLC-kNN classi?er, followed by the MLC-GIS and CB-GIS classi?ers. According to the best micro-averaged F1 scores, the new methods (CB-GIS, MLC-CIS, MLC-kNN) have improvements of 4.48%, 4.65% and 4.76% over kNN, 1.84%, 1.92% and 2.12% over the centroidbased and 5.26%, 5.34% and 5.45% over GIS respectively. With respect to the best macro-averaged F1 scores, the new methods (CB-GIS, MLC-CIS, MLC-kNN) have improvements of 10.29%, 10.19% and 10.45% over kNN, respectively, 0.1%, 0.03% and 0.29% over the centroid-based and 3.75%, 3.68% and 3.94% over GIS respectively.