Data Mining Feature Clustering Algorithm in Text Classification

Yaganteeswarudu, A[1]*., Siva Prasad, K[2]., Vishnuvardhan, Y[3] and Chandra Sekhar., V[1]

1. Department of CSE, SJCET, Kurnool, A.P, INDIA
2. Department of CSE, AVR & SVRCET, Kurnool, A.P., INDIA
3. Department of CSE, NITK, Karnataka, INDIA

KEYWORDS

Fuzzy similarity, feature clustering, feature extraction, feature reduction, text classification

Abstract: *Feature clustering is a powerful method to reduce the dimensionality of feature vectors for text classification. In this paper, we propose a fuzzy similarity-based self-constructing algorithm for feature clustering. The words in the feature vector of a document set are grouped into clusters, based on similarity test. Words that are similar to each other are grouped into the same cluster. Each cluster is characterized by a membership function with statistical mean and deviation. When all the words have been fed in, a desired number of clusters are formed automatically. We then have one extracted feature for each cluster. The extracted feature, corresponding to a cluster, is a weighted combination of the words contained in the cluster. By this algorithm, the derived membership functions match closely with and describe properly the real distribution of the training data. Besides, the user need not specify the number of extracted features in advance, and trial-and-error for determining the appropriate number of extracted features can then be avoided. Experimental results show that our method can run faster and obtain better extracted features than other methods.*