Improved normalization and standardization techniques for higher purity in K-means clustering

ABSTRACT

Clustering is basically one of the major sources of primary data mining tools, which make researchers understand the natural grouping of attributes in datasets. Clustering is an unsupervised classification method with aim of partitioning, where objects in the same cluster are similar, and objects belong to different clusters vary significantly, with respect to their attributes. The K-means algorithm is a famous and fast technique in non-hierarchical cluster algorithms. Based on its simplicity, the K-means algorithm has been used in many fields. This paper proposes improved normalization and standardization techniques for higher purity in K-means clustering experimented with benchmark datasets from UCI machine learning repository and it was found that all the proposed techniques' performance was much higher compared to the conventional K-means and the three classic transformations, and it is evidently shown by purity and Rand index accuracy results.

Keyword: Normalization; Standardization; K-means algorithm; Clustering; Purity; Rand index