World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:11, No:06, 2017

Clustering Categorical Data Using the K-Means Algorithm and the Attribute's Relative Frequency

Authors : Semeh Ben Salem, Sami Naouali, Moetez Sallami

Abstract : Clustering is a well known data mining technique used in pattern recognition and information retrieval. The initial dataset to be clustered can either contain categorical or numeric data. Each type of data has its own specific clustering algorithm. In this context, two algorithms are proposed: the k-means for clustering numeric datasets and the k-modes for categorical datasets. The main encountered problem in data mining applications is clustering categorical dataset so relevant in the datasets. One main issue to achieve the clustering process on categorical values is to transform the categorical attributes into numeric measures and directly apply the k-means algorithm instead the k-modes. In this paper, it is proposed to experiment an approach based on the previous issue by transforming the categorical values into numeric ones using the relative frequency of each modality in the attributes. The proposed approach is compared with a previously method based on transforming the categorical datasets into binary values. The scalability and accuracy of the two methods are experimented. The obtained results show that our proposed method outperforms the binary method in all cases.

Keywords : clustering, unsupervised learning, pattern recognition, categorical datasets, knowledge discovery, k-means
Conference Title : ICMLA 2017 : International Conference on Machine Learning and Applications
Conference Location : Copenhagen, Denmark
Conference Dates : June 11-12, 2017