Article Abstract

Title:	**A hybridized K-means clustering approach for high dimensional dataset**
Author(s):	Rajashree Dash[1], Debahuti Mishra[*], Amiya Kumar Rath[2], Milu Acharya[3]
Address(es):	[1,*,3]Institute of Technical Education and Research, Bhubaneswar, Orissa, INDIA [2] Director, College of Engineering Bhubaneswar, Orissa, INDIA [*]Corresponding Authors' e-mails: debahuti@iter.ac.in, rajashree_dash@yahoo.co.in ,amiyamiya@rediffmail.com, milu_acharya@yahoo.com
Journal:	*International Journal of Engineering, Science and Technology*, Vol. 2, No. 2, 2010, pp. 59-66.
Abstract:	Due to incredible growth of high dimensional dataset, conventional data base querying methods are inadequate to extract useful information, so researchers nowadays is forced to develop new techniques to meet the raised requirements. Such large expression data gives rise to a number of new computational challenges not only due to the increase in number of data objects but also due to the increase in number of features/attributes. Hence, to improve the efficiency and accuracy of mining task on high dimensional data, the data must be preprocessed by an efficient dimensionality reduction method. Recently cluster analysis is a popularly used data analysis method in number of areas. K-means is a well known partitioning based clustering technique that attempts to find a user specified number of clusters represented by their centroids. But its output is quite sensitive to initial positions of cluster centers. Again, the number of distance calculations increases exponentially with the increase of the dimensionality of the data. Hence, in this paper we proposed to use the Principal Component Analysis (PCA) method as a first phase for K-means clustering which will simplify the analysis and visualization of multi dimensional data set. Here also, we have proposed a new method to find the initial centroids to make the algorithm more effective and efficient. By comparing the result of original and new approach, it was found that the results obtained are more accurate, easy to understand and above all the time taken to process the data was substantially reduced.
Keywords:	Cluster analysis, K-means Algorithm, Dimensionality Reduction, Principal Component Analysis, Hybridized K-means algorithm