Abstract

The problems that arise in the data when the data consists of numeric attributes and categories. To handle such problems required special techniques to perform clustering on the data made by numerical and categories attributes. One clustering algorithm used to perform clustering on a mixture of numerical data and the category is Semi-supervised Regression Model. Clustering process is done by combining multiple linear regression for numerical data and k-modes clustering for categorical data. Data are grouped according to the smallest value of the least square error for numeric attributes and dissimilarity measures for the attribute category of the center of a cluster. The result is a semi-supervised regression algorithm model suitable for application at the data that has numeric attributes and categories where the data is the range of values on numeric attributes is not too far away and a small standard deviation value and the data categories do not have an equal value distribution for different clusters.

Keyword : clustering, semi-supervised, k-mode, multiple regression, least square error, dissimilarity measure