ABSTRAK

Clustering merupakan suatu cara pengelompokan data di *data mining*. Metode *clustering* yang sering di gunakan adalah *K-Means*. Masalah yang akhir-akhir ini sering timbul adalah ketika data yang akan diklasterisasi adalah data kategori, tidak memungkinkan bila kita juga menerapkan metode *K-Means* untuk mengklaster data kategori. Data kategori merupakan data yang diambil dari suatu himpunan nilai tertentu, yang tidak harus berupa angka atau pecahan misalkan warna mata, warna kulit, dan nama negara. Data kategori tidak dapat diukur dan diurutkan karena tidak dapat dibandingkan mana yang memiliki nilai lebih besar dan nilai yang lebih kecil. Konsep ukuran jarak pada *clustering* data kategori berbeda dengan *clustering* data numerik yang menggunakan *K-Means*. Maka dari itu dikembangkannya *K-Modes* yaitu hasil pengembangan *K-Means* yang merupakan metode pengklasteran dalam mengelompokkan tipe data kategori. Pada *K-Modes* Konvensional dimana $x_j = y_j$ tidak lagi bernilai 0 tetapi $1-w_{ij}$. Nilai w_{ij} adalah perkalian perbandingan nilai atribut di *cluster* dengan perbandingan nilai atribut di dataset. Hal ini membuat dalam menghasilkan pembentukan *cluster*nya lebih rinci lagi, sehingga kesamaan intra *cluster* bertambah kuat. Penelitian ini bertujuan untuk mengimplementasikan algoritma *clustering K-Modes* menggunakan *New Dissimilarity Measure* pada beberapa data kategori. Nilai *purity* yang dihasilkan oleh *clustering K-Modes* menggunakan *New Dissimilarity Measure* dapat mencapai 0.76, sedangkan pada *K-Modes* Konvensional 0.61. Sedangkan pada evaluasi menggunakan *F-Measure* pada *K-Modes New Dissimilarity Measure* menghasilkan nilai *F-Measure* sebesar 0.80.

ABSTRACT

Clustering is a way of grouping data in data mining. Clustering method which is often used is K-Means. Lately problem often arises when the data that will be clustered is categorical data, it is not possible if we also apply the K-Means method. Categorical data is a data that take from a certain set of values, which do not have to be an integer or fractional, for example eye color, skin color, and country names. Categorical data can not be measured and sorted as it can not be compared which one has a greater value and a smaller value. Distance measurement concept in categorical data clustering is different with numerical data clustering that used K-Means. Thus the K-Modes developed which is the development from K-Means which is the clustering method to classify categorical data type. In the Conventional K-Modes where xj = yj no longer be 0 but 1-wij. Wij is the multiplication of attribute value comparison in cluster with attribute value comparison in dataset. It made generating cluster more detail, so that the intra-cluster similarity become better. This study aims to implement the K-Modes clustering algorithm using New Dissimilarity Measure in several categorical data. Purity value produced by the K-Modes clustering using New Dissimilarity Measure can reach 0.76, while the Conventional K-Modes only 0.61. While on the evaluation using F-Measure on K-Modes new generates F-Measure values of 0.80.