ANALISIS CLUSTER UNTUK DATA CAMPURAN KATEGORIK DAN NUMERIK

(Cluster Analysis for Mixed Cagtegorical and Numeric Data Types)

Yuliani Setia Dewi
Jurusan Matematika FMIPA Universitas Jember

Abstract: Various clustering algorithms have been developed to group data into clusters. This paper describes clustering objects with mixed categorical and numeric data types. The methods used are two step clustering and transforms mixed data using nonlinear principal component analysis then groups the output resulted using hierarchical aglomerative clustering. The results show that the number of optimal cluster using both methods have the same optimal number of cluster but the rank of ratios of distance measure and distribution of cluster membership are different.

Keywords: Two step clustering, nonlinear principal component analysis, mixed data, transform.

I. PENDAHULUAN

Tujuan dari analisis cluster adalah mengelompokkan objek-objek berdasarkan kesamaan karakteristik di antara objek-objek tersebut. Objek tersebut akan diklasifikasikan ke dalam satu atau lebih *cluster* (kelompok) sehingga objek-objek yang mempunyai kesamaan yang tinggi akan berada dalam satu *cluster* dan objek-objek yang mempunyai ketidaksamaan yang tinggi akan berada pada cluster yang berbeda (Han & Kamber, 2001).

Banyak metode yang dapat digunakan untuk melakukan pengclusteran, yang dapat dikategorikan sebagai metode berhirarki (*Hierarchical Methods*) dan metode tidak berhirarki (*Nonhierarchical Methods*). Metode berhirarki terbagi menjadi dua, yaitu metode *agglomerative* (penggabungan) dan metode *divisive* (pemecahan). Pada metode berhirarki penggabungan objek ke dalam kelompok-kelompok dilakukan dengan menggunakan tiga metode, yaitu metode pautan tunggal (*Single Linkage Method*), metode pautan lengkap (*Complete Linkage Method*) dan metode rata-rata kelompok (*Average Linkage Method*).

Untuk mengelompokkan objek-objek kedalam satu kelompok digunakan ukuran kemiripan atau ketidakmiripan antar objek atau cluster yang diukur dengan menggunakan ukuran jarak. Telah banyak dikenal ukuran jarak untuk variabel yang mempunyai tipe sama. Jarak Euclid sering digunakan ketika variabel-variabelnya bertipe numerik. Jarak