Abstrak

Clustering atau pengelompokan merupakan salah satu metode yang digunakan untuk memecahkan persoalan pada *data mining*. Terdapat beberapa algoritma yang digunakan untuk menyelasikan permasalahan pada *clustering*, salah satunya yaitu *k-means clustering*. Namun terdapat beberapa kelemahan pada algoritma *k-means* sehingga dapat mengurangi akurasi dan efisiensi. Salah satu kelemahan tersebut yaitu penentuan pusat *cluster* atau inisialisasi *centroid* secara acak dapat menyebabkan hasil *cluster* yang tidak akurat karena pusat *cluster* yang tidak tersebar ke seluruh *dataset*. Pada penelitian ini akan dibahas mengenai metode untuk meminimalisir kelemahan tersebut sehingga dapat menghasilkan akurasi dan efisiensi yang lebih baik. Penentuan pusat *cluster* atau inisialisasi *centroid* mengunakan algoritma *k-means++* menyebabkan persebaran *cluster* pada *data point* lebih merata jika dibandingkan dengan menggunakan *k-means* yang memilih secara acak pusat-pusat *cluster* yang dituju. Metode *canopy* atau *canopy clustering* ialah suatu metode yang biasa digunakan sebagai langkah awal dari metode *clustering* seperti *k-means* dan *expectation maximization* (EM). Metode *canopy* memiliki algoritma yang sederhana, cepat, dan akurat untuk mengelompokkan data ke dalam kelompok tertentu. Dengan mengkombinasikan metode *canopy* dengan metode *clustering* lainnya seperti *k-means* akan mengurangi penghitungan jarak antara *data point* yang biasanya memakan waktu dengan cara membentuk *canopy* untuk membatasi jumlah *data point* yang akan dihitung masing-masing jaraknya. Dengan menggunakan metode *canopy* dan *k-means++* tingkat akurasi yang dihasilkan yaitu 98.3% untuk *dataset* dim1024 dan 87.7% untuk dataset Iris. Sedangkan dari segi waktu *canopy* dan *k-means++* memperoleh hasil dalam waktu 7 detik pada *dataset* dim1024 dan 0.77 detik pada *dataset* iris.

Kata kunci: *Data mining, Clustering, K-means Clustering, Canopy Clustering, Centroid Initialization. K-Means++.*