JURUSAN TEKNIK INFORMATIKA

FAKULTAS TEKNOLOGI INFORMASI

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

USULAN TUGAS AKHIR

1. IDENTITAS PENGUSUL

Nama : Eric Budiman Gosno

NRP : 5109100153

Dosen Wali : Waskitho Wibisono, S.Kom, M.Eng., Ph.D.

2. JUDUL TUGAS AKHIR

"Implementasi K-Dimensional Tree untuk Proses Inisialisasi pada Algoritma K-Means Clustering"

"Implementation of K- Dimensional Tree for Initialising K-Means Clustering Algorithm"

3. URAIAN SINGKAT

Clustering merupakan salah satu dasar untuk menyelesaikan permasalahan penggalian data, *machine learning*, statistik, dan pengenalan pola. Salah satu metode *clustering* dasar yang telah dikenal, dan diaplikasikan secara luas adalah *K-Means Clustering*.

K-Means Clustering merupakan metode optimisasi local yang sensitif terhadap pemilihan posisi awal dari titik tengah *cluster* sehingga pemilihan posisi awal dari titik tengah/*seeds* cluster yang buruk akan mengakibatkan algoritma *K-Means Clustering* terjebak dalam *local optimization*.

Pada tugas akhir ini penulis mencoba untuk melakukan optimasi pada algoritma Katsavoudinis dengan menggunakan struktur data *K-Dimensional Tree* untuk proses pemilihan K seed awal dari cluster pada algoritma *K-Means Clustering*.

4. PENDAHULUAN

4.1 LATAR BELAKANG

Clustering merupakan salah satu dasar untuk menyelesaikan permasalahan penggalian data, *machine learning*, statistik, dan pengenalan pola. Secara umum, Permasalahan *Clustering* dapat dibagi menjadi 2 jenis yaitu *Hierarchial Clustering*, dan *Partitional Clustering*.

Partitional Clustering merupakan suatu permasalahan untuk membagi/melakukan partisi dari sebuah dataset ke dalam sejumlah kelompok data yang disebut sebagai *cluster* dengan memaksimalkan total nilai kemiripan/*similiarity* antar data-data yang ada pada *cluster* yang sama. Mencari hasil partisi yang menghasilkan nilai optimal pada *Partitional Clustering* merupakan permasalahan *NP-Complete*, tetapi terdapat beberapa strategi suboptimal yang dapat memberikan solusi yang kompetitif dalam kompleksitas linear. Salah satu strategi suboptimal *partitional clustering* yang telah dikenal, dan diaplikasikan secara luas adalah *K-Means Clustering*.

K-Means Clustering merupakan metode optimisasi local yang sensitif terhadap pemilihan posisi awal dari titik tengah *cluster* sehingga pemilihan posisi awal dari titik tengah/*seeds* cluster yang buruk akan mengakibatkan algoritma *K-Means Clustering* terjebak dalam *local optimization* [1]. Saat ini telah terdapat beberapa metode yang telah diusulkan untuk proses inisialisasi dari *K-Means Clustering* seperti *Forgy Algorithm(FA)* [2,3], *MacQueen Aproach(MA)* [4], *Binary Splitting(BS)* [5], *KKZ/Katsavoudinis Algorithm* [6], hingga yang terakhir *Cluster Centre Initialisation Method(CCIA)* [7].

Pada tugas akhir ini penulis mencoba untuk melakukan optimasi pada *Katsavoudinis Algorithm* dengan menggunakan struktur data *K-Dimensional Tree* untuk proses pemilihan K seed awal dari cluster pada algoritma *K-Means Clustering*.

4.2 RUMUSAN MASALAH

Beberapa permasalahan yang dibahas dalam tugas akhir ini dapat diuraikan sebagai berikut:

1. Bagaimana mengimplementasikan struktur data *K-Dimensional Tree* pada proses inisialisasi pemilihan titik tengah *cluster* dari algoritma *K-Means Clustering*?

2. Bagiamana performa dari algoritma *K-Means Clustering* yang menggunakan struktur data *K-Dimensional Tree* dibandingkan dengan metode inisialisasi titik tengah *cluster* sebelumnya?

4.3 BATASAN MASALAH

Permasalahan yang dibahas dalam tugas akhir ini memiliki batasan – batasan sebagai berikut:

1. Implementasi dari metode *K-Means Clustering*, dan stuktur data *K-Dimensional Tree* menggunakan bahasa pemrograman C# dengan bantuan perangkat lunak Microsoft Visual Studio 2010
2. Pustaka yang digunakan saat mengimplementasi algoritma menggunakan pustaka yang terdapat dalam *.Net Framework Class Library*
3. Performa yang diteliti meliputi waktu eksekusi(*running time)* dari program dan nilai *Distortion* minimum dari hasil *clustering* yang dihasilkan

4.4 TUJUAN TUGAS AKHIR

Tujuan dari pengerjaan tugas akhir ini adalah mengimplementasikan, dan melakukan uji performa dari algoritma *K-Means Clustering* dengan struktur data *K-Dimensional Tree* pada proses inisialisasi titik tengah.

4.5 MANFAAT TUGAS AKHIR

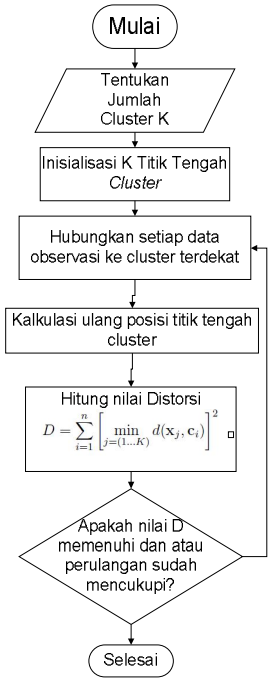
Manfaat yang diharapkan dari hasil tugas akhir ini adalah memberikan referensi untuk metode inisialisasi titik tengah cluster untuk algoritma *K-Means Clustering* yang dapat menghasilkan performa yang baik dalam segi waktu, kompleksitas, dan akurasi hasil *clustering*.

5 TINJAUAN PUSTAKA

5.1 Algoritma K-Means Clustering

Algoritma K-Means Clustering merupakan salah satu metode dari analisa cluster/*clustering* yang bertujuan untuk melakukan partisi dari sebuah dataset ke dalam *k* cluster. Dalam prosesnya, K-Means Clustering akan mencari titik tengah/*seeds* dari semua *cluster* yang menghasilkan total nilai jarak/*distortion* dari setiap data observasi ke titik tengah

cluster seminimal mungkin. Alur metode dari algoritma *K-Means Clustering* dapat dilihat pada Gambar 1.



Gambar 1 Alur Metode K-Means Clustering

Algoritma *K-Means Clustering* dengan nilai *K* tertentu secara umum memiliki kompleksitas *NP-Hard* [8]. Tetapi, jika diberikan nilai *D/*Distorsi yang diharapkan, maka Algoritma *K-Means Clustering* dapat diselesaikan dengan kompleksitas $O(n^{dk+1} \log n)$ dimana *n* menyatakan banyaknya data observasi [9]. Keunggulan dari K-Means Clustering adalah mudah untuk diimplementasi, dan diaplikasikan pada dataset yang berukuran besar seperti pada permasalahan segmentasi pasar, visi computer, *geostatic*, astronomi, dan pertanian.

5.2 Algoritma KKZ/Katsavounidis

Algoritma KKZ/Katsavounidis adalah metode inisialisasi titik tengah/*seed* dari *cluster* pada algoritma *K-Means Clustering*. Algoritma ini dimulai dengan memilih sebuah data *x* (diutamakan data memiliki posisi ujung/*edge* pada dataset) sebagai *seed* pertama. Algoritma kemudian akan mencari data yang memiliki jarak terjauh atau nilai kemiripan terkecil dengan *seed* pertama, dan datatersebut akan ditandai sebagai *seed* kedua. Algoritma akan terus

mencari data yang memiliki jarak paling jauh dengan *seed* terdekat hingga telah didapatkan sebanyak *K seed* sebagai inisialisasi titik tengah bagi algoritma *K-Means Clustering*.

Dalam setiap iterasi untuk mencari *seed* baru algoritma KKZ akan melakukan $0.5 \times N\ X\ (N-1)$ kalkulasi perhitungan jarak antara 2 data sehingga dapat dikatakan bahwa algoritma KKZ memiliki kompleksitas O(*MK*) dimana *K* menyatakan banyaknya cluster dan *M* menyatakan banyaknya kalkulasi perhitungan jarak antara 2 data (M =$0.5 \times N\ X\ (N-1)$). Nilai kompleksitas ini sama dengan 1 kali iterasi *Lloyd-Forgy* yang merupakan metode paling dasar dari proses inisialisasi *K-Means Clustering*.

Algoritma KKZ memiliki kelemahan dalam menangani dataset yang memiliki banyak *noise* dan *outlier*. Hal ini disebabkan karena algoritma KKZ akan cenderung untuk memilih data *noise* yang memiliki parameter ekstrim sebagai *seed* dari *cluster* yang tentu akan mengakibatkan naiknya nilai *Distortion* pada proses *K-Means Clustering* [10].

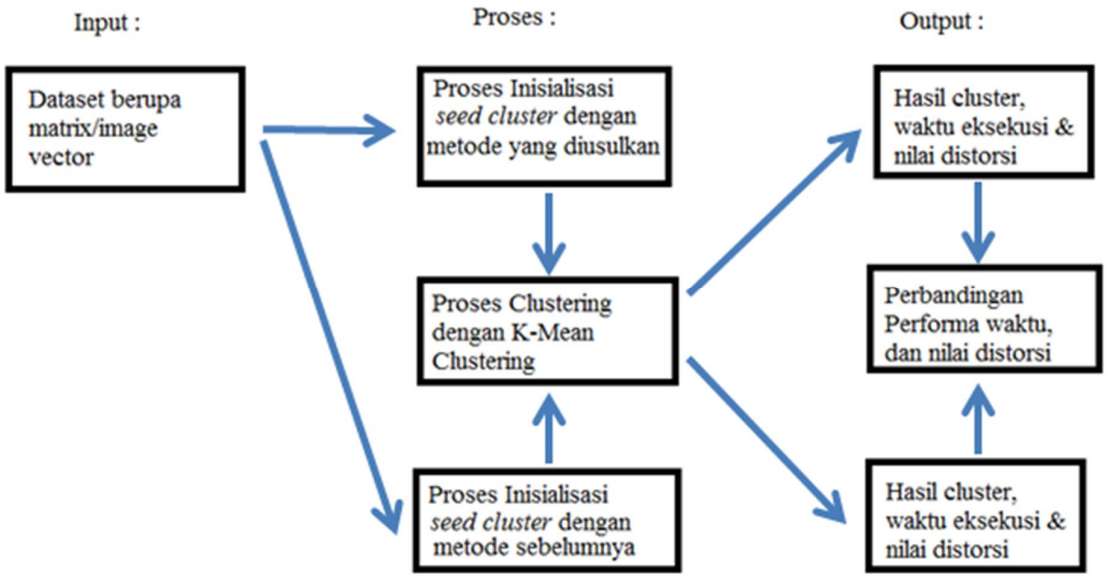
5.3 K-Dimensional Tree

K-Dimensional Tree (*KD-Tree*) [11] adalah data struktur yang bersifat *space-partitioning*, dan merupakan kasus special dari *binary space partitioning tree*. *KD-Tree* bertujuan untuk menngatur poin-poin dalam *k–dimensional space*. *KD-Trees* umumnya diaplikasikan dalam pencarian yang memiliki banyak kunci pencarian(*multidimensional key)* seperti *range search*, dan *nearest neighbor search*.

Dalam implementasi, *KD-Tree* adalah *binary tree* dimana setiap *node* pada *binary tree* tersebut adalah sebuah *point* berdimensi *k*. Setiap *node* yang bukan merupakan *leaf* pada *KD-Tree* akan menghasilkan sebuah *hyperplane* yang memisakan sebuah ruang menjadi 2 bagian. Setiap poin yang berada di daerah sebelah kiri *hyperplane* merepresentasikan *node* yang berada di *subtree* sebelah kiri. Demikian pula dengan setiap poin yang berada di daerah sebelah kanan *hyperplane* akan merepresentasikan *node* yang berada di *subtree* sebelah kanan.

6 METODOLOGI

Berikut merupakan alur, input dan output dari algoritma yang akan dibuat seperti pada gambar 3.



Gambar 3: Gambaran Alur Kerja Tugas Akhir

Keterangan gambar 3 :

1. Input dari aplikasi ini adalah dataset berupa matrix/image vector. Terdapat 2 jenis input yang akan digunakan yaitu dataset sintetis/buatan dan dataset dunia nyata(*real-world dataset)*. Untuk dataset sintetis, akan digunakan *syntethic multivariate Gaussian data*. Sedangkan untuk *real-world dataset* akan digunakan *Pen-Based Recognition of Handwritten Digits Database*, dan *Image Segmentation Database*.
2. Input akan dimasukkan ke dalam aplikasi *K-Means Clustering* yang menggunakan proses inisialisasi *seed cluster* yang diusulkan, dan aplikasi yang menggunakan proses inisialisasi *seed cluster* sebelumnya.
3. Akan dilakukan perbandingan performa, dan hasil dari output yang berasal proses inisialisasi *seed cluster* yang diusulkan dengan hasil dari proses inisialisasi *seed cluster* sebelunnya.

6.1 Langkah Pengerjaan

Dalam pembuatan aplikasi ini, penulis telah menyusun beberapa langkah dalam pengerjaan aplikasi. Berikut langkah kerja yang telah disusun penulis :

6.1.1 Studi Literatur

Studi literatur mencakup pembelajaran pada algoritma-algoritma terkait, analisa *prove of correctness*, serta perhitungan kompleksitas. Pada tahap ini juga dilakukan studi terhadap beberapa implementasi terdahulu apabila tersedia untuk dijadikan acuan pada tahap selanjutnya.

6.1.2 Implementasi Algoritma dan Uji Coba

Hasil studi pada tahap sebelumnya menjadi dasar pada tahap implementasi. Bahasa yang digunakan adalah C# dengan bantuan IDE Microsoft Visual Studio 2010. Diharapkan pada tahap ini didapatkan bentuk implementasi yang cukup optimal pada algoritma *K-Means*

Clustering beserta dengan metode inisialisasi *seed cluster* yang digunakan agar didapatkan hasil yang relevan pada tahap eksperimen.

6.1.3 Eksperimen dan Evaluasi

Pada tahap ini hasil implementasi dari *K-Means Clustering* dengan beberapa jenis metode inisialisasi *seed cluster* akan diujikan pada dataset yang telah ditentukan. Performa, dan hasil dari proses *Clustering* antara metode yang diusulkan dengan metode sebelumnya akan menjadi dasar acuan dari tahap evaluasi, dan pengambilan kesimpulan

6.1.4 Penyusunan buku tugas akhir

Tahap ini merupakan penyusunan laporan berupa buku tugas akhir sebagai dokumentasi pelaksanaan tugas akhir, yang mencakup seluruh teori, implementasi, serta hasil pengujian yang telah dikerjakan.

7 JADWAL PEMBUATAN TUGAS AKHIR

Tugas akhir ini diharapkan bisa dikerjakan sesuai jadwal, sebagai berikut.

Tahapan	2013																			
	Maret				April				Mei				Juni				Juli			
Penyusunan Proposal																				
Studi Literatur																				
Implementasi																				
Pengujian dan Evaluasi																				
Penyusunan Buku																				

8 DAFTAR PUSTAKA

[1] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data clustering: A review," *ACM Computation Surveys*, vol. 31, no. 3., pp. 264–323, 1999.

[2] M.R. Anderberg, *Cluster Analysis for Applications*. New York, United States: Academic Press, 1973.

[3] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, vol. 21, pp. 768–769, 1965.

[4] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[5] Y. Linde, A. Buzo, and R.M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. 28, pp. 84–94, 1980.

[6] I. Katsavounidis, C.C.J. Kuo, and Z. Zhen, "A new initialization technique for generalized lloyd iteration," *IEEE Signal Processing Letter*, vol. 1, no. 10, pp. 144–146, 1994.

[7] Shehroz S. Khana and Amir Ahmadb, "Cluster center initialization algorithm for K-means clustering," *Pattern Recognition Letters*, vol. 25, no. 11, pp. 1293–1302, August 2004.

[8] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The Planar k-Means Problem is NP-Hard," *Lecture*

Notes in Computer Science, vol. 5431, pp. 274–285, 2009.

[9] M. Inaba, N. Katoh, and H. Imai, "Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering," in *Proceedings of 10th ACM Symposium on Computational Geometry*, 1994, pp. 332–339.

[10] Stephen J Redmond and Conor Heneghan, "A method for initialising the K-means clustering algorithm using kd-trees," *Pattern Recognition Letters*, vol. 28, no. 8, pp. 965–973, June 2007.

[11] J.L Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509-517, 1975.