JURUSAN TEKNIK INFORMATIKA

FAKULTAS TEKNOLOGI INFORMASI

INSTITUT TEKNLOGI SEPULUH NOPEMBER

USULAN TUGAS AKHIR

1. IDENTITAS PENGUSUL

NAMA	**:**	**Ratih Kirana Diantari**
NRP	**:**	**5110100051**
DOSEN WALI	**:**	**Dr. Ir. Siti Rochimah, M.T.**
DOSEN PEMBIMBING	**:**	**1. Dr. Chastine Fatichah, S.Kom, M.Kom.**
		2. Rully Soelaiman, S.Kom, M.Kom.

2. JUDUL TUGAS AKHIR

"Identifikasi Parameter yang Berpengaruh pada Kinerja Algoritma Clustering dengan Metode Evidence Accumulation"

3. LATAR BELAKANG

Clustering adalah metode penganalisaan data yang bertujuan untuk mengelompokkan data dengan karakteristik yang sama ke suatu "wilayah" yang sama dan data dengan karakteristik yang berbeda ke "wilayah" yang lain. Ada dua pendekatan utama dalam metode *clustering* yaitu pendekatan partisi (*Partition-based Clustering*) dan pendekatan hierarki (*Hierarchical Clustering*). Salah satu contoh untuk pendekatan partisi adalah *Consensus Clustering*.

Consensus Clustering menggabungkan beberapa *clustering* tanpa akses ke fitur-fitur yang mendasari data, sehingga menghasilkan *clustering* akhir yang kuat dibandingkan dengan beberapa *clustering* lainnya. Algoritma *consensus clustering* menghasilkan *clustering* yang lebih baik, menemukan pengelompokan *clustering* yang tidak terjangkau oleh algoritma *clustering* tunggal, kurang sensitif terhadap *noise*, *outlier* atu variasi sampel; dan mampu mengintegrasikan solusi dari berbagai sumber data atau atribut. *Consensus Clustering* juga dapat berguna dalam berbagai domain. Misalnya, mengelompokkan data kategori dapat dianggap sebagai masalah *consensus clustering* dimana masing-masing fitur diskrit dipandang sebagai pengelompokan sederhana dari data [**1**]. Kelemahan dari *consensus clustering* ini adalah kesamaan berpasangan seringkali tidak menggambarkan kesamaan ukuran yang baik antara titik data, terutama ketika jumlah basis-*clustering* terbatas [**2**]. Sebagian besar algoritma ini kembali ke *consensus clustering* tunggal sebagai hasil akhir. Ada dua pendekatan

---

utama pada solusi *consensus*, yaitu berpatokan pada model probabilistik dan mengukur kesamaan antara dua *clustering*.

Pada Tugas Akhir ini, *Consensus Clustering* diolah menggunakan metode *Evidence Accumulation* (EA). Metode ini mampu menangani partisi lengkap dalam *ensemble* serta parsial. Ada beberapa paramater yang diidentifikasi, misalnya nilai *k* dan *treshold* dimana beberapa parameter ini akan mempengaruhi kinerja *clustering*.

4. RUMUSAN MASALAH

Rumusan masalah yang diangkat dalam Tugas Akhir ini adalah sebagai berikut:

a. Memahami konsep *consensus clustering*.
b. Mengimplementasi paramater metode *Evidence Accumulation* pada *consensus clustering*.
c. Menyusun ujicoba menggunakan metode *Evidence Accumulation* pada *consensus clustering*.

5. BATASAN MASALAH

Adapun batasan ruang lingkup permasalahan dari pengerjaan Tugas Akhir ini adalah sebagai berikut:

a. Implementasi menggunakan perangkat lunak MATLAB.
b. Data set yang digunakan adalah Ecoli, Glass, Iris, Seeds, Vowel, dan Wine yang didapatkan dari http://archive.ics.uci.edu/ml/. Keenam data set tersebut dipilih karena memiliki jumlah atribut dan jumlah data yang beragam.Untuk data Iris, Wine, Glass, dan Ecoli tidak memiliki *missing value*.

6. TUJUAN PEMBUATAN TUGAS AKHIR

Tujuan dari pembuatan Tugas Akhir ini adalah sebagai berikut:

a. Mengidentifikasi paramater metode *Evidence Accumulation* untuk menyelesaikan permasalahan *consensus clustering*.
b. Mengevaluasi kinerja metode *Evidence Accumulation* dengan melakukan ujicoba.

7. MANFAAT TUGAS AKHIR

Manfaat yang diharapkan dari Tugas Akhir ini adalah untuk mengetahui hasil yang terbaik untuk *Consensus Clustering*.

8. TINJAUAN PUSTAKA

a. *Evidence Accumulation* (EA)
   Ide *Evidence Accumulation* adalah menggabungkan hasil dari beberapa *clustering*. Pertama, *clustering ensemble* – satu set partisi objek dibentuk. Cara yang berbeda untuk menghasilkan partisi data adalah menerapkan algoritma *clustering* yang berbeda, dan menerapkan algoritma *clustering* yang sama dengan nilai parameter atau inisialisasi yang berbeda. Selanjutnya kombinasi dari representasi data yang berbeda dan algoritma *clustering* juga dapat

memberikan banyak data partisi yang berbeda secara signifikan [**3**]. Algoritma ini memiliki konsep bahwa setiap partisi dipandang sebagai bukti independen organisasi data, partisi data individu yang digabungkan, berdasarkan mekanisme *voting*, untuk menghasilkan nxn kesamaan matriks baru antara pola n. Kentungan algoritma ini dapat menghindari masalah korespondensi label, yang mempengaruhi skema pengelompokan *ensemble* lainnya. EA memiliki pola dimensi n-d, strategi yang diusulkan mengikuti pendekatan *split* dan *merge*:

Split : Mengurangi data multidimensi menjadi beberapa *cluster* kecil. Algoritma K-Means melakukan pengurangan ini, dengan berbagai hasil *cluster* diperoleh inisialisasi acak algoritma.

Combine : Mengatasi partisi dengan jumlah *cluste*r yang berbeda, menggunakan "*voting*" untuk menggabungkan hasil *clustering*, sesuai ke ukuran baru dari kesamaan antara pola. Asumsinya adalah pola-pola *cluster* yang alami sangat mungkin menjadi "letak" di *cluster* yang sama di *clustering* yang berbeda. Mengambil *co-occurences* dalam pola berpasangan di *cluster* yang sama sebagai hasil untuk asosiasi mereka, partisi data dijalankan dalam K-Means. Mengambil pola berpasangan *co-occurence* dalam *cluster* yang sama sebagai "*voting*" untuk asosiasi, data partisi yang dihasilkan oleh K-Means dipetakan pada matriks *co-association* nxn, yang ditunjukkan oleh Persamaan 1:

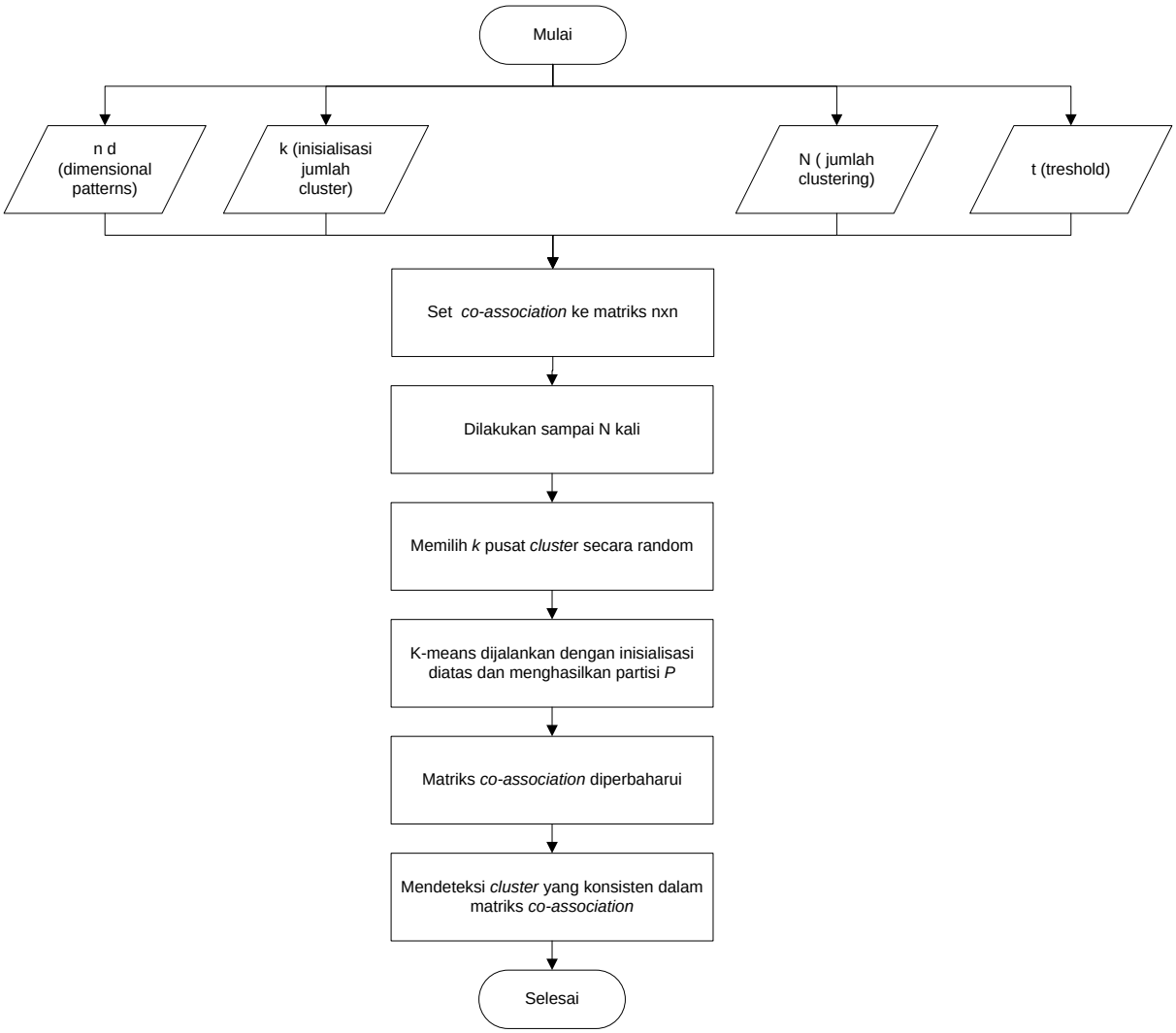
$$co_{assoc\,(i,j)} = \frac{votes_{ij}}{N} \quad (1)$$

dimana N merupakan jumlah *clustering* dan $votes_{ij}$ merupakan berapa kali pasangan berpola (i,j) ditugaskan untuk *cluster* yang sama antara *clustering* N.

Merge : Untuk membentuk *cluster* ke bentuk semula, menggunakan algoritma *Minimum Spanning Tree* (MST). Yaitu, memotong rantai yang lemah pada ambang *t*, sehingga penggabungan *cluster* yang dihasilkan dalam tahap pemisahan [**4**].

9. RINGKASAN ISI TUGAS AKHIR

Consensus Clustering membahas masalah meringkas satu set *clustering* yang diperoleh untuk data set tertentu menjadi partisi *consensus* tunggal. Ada beberapa pendekatan yang diusulkan, salah satunya adalah *Evidence Accumulation* (EA). Metode ini memiliki pendekatan *split* dan *merge*, dimana memiliki 2 parameter yaitu k-jumlah *cluster* untuk algoritma K-Means dan t sebagai *treshold* pada MST. Untuk mendeteksi *cluster* yang konsisten menggunakan teknik *single-link* (SL). Teknik ini mencari *voting* terbanyak untuk setiap pasangan pola, kemudian menggabungkan pola dalam *cluster* yang sama. Apabila pola terbentuk dalam *cluster* yang berbeda sebelumnya, maka akan bergabung dengan kelompoknya. Untuk setiap pola yang tersisa yang tidak termasuk dalam *cluster* membentuk *cluster* tunggal [**4**]. Secara garis besar, proses yang dilakukan *Evidence Accumulation* ditunjukkan pada Gambar 1.



Gambar 1. Flowchart metode EA

10. METODOLOGI

a. Penyusunan proposal tugas akhir

Proposal Tugas Akhir ditulis untuk mengajukan ide atas pengerjaan Tugas Akhir. Proposal ini juga mengandung proyeksi dari ide Tugas Akhir yang diajukan.

b. Studi literatur

Pada proses ini dilakukan studi lebih lanjut terhadap konsep-konsep yang terdapat pada jurnal, buku, artikel, dan literatur yang menunjang. Studi dilakukan untuk mendalami konsep algoritma EA untuk menyelesaikan permasalahan yang muncul pada proses pengerjaan Tugas Akhir ini.

c. Implementasi algoritma

Implementasi merupakan tahapan untuk membangun sistem tersebut. Algoritma yang akan diimplementasikan yaitu algoritma clustering dengan metode *Evidence Accumulation*. Implementasi diproses menggunakan MATLAB.

d. Pengujian dan evaluasi

Pada tahap ini dilakukan uji coba dengan menggunakan beberapa data set menggunakan metode *Evidence Accumulation*.

e. Penyusunan Buku Tugas Akhir

Pada tahap ini dilakukan penyusunan laporan yang menjelaskan dasar teori dan metode yang digunakan dalam Tugas Akhir ini serta hasil yang telah dikerjakan. Sistematika penulisan buku Tugas Akhir secara garis besar antara lain:

1. Pendahuluan
    a. Latar Belakang
    b. Rumusan Masalah
    c. Batasan Tugas Akhir
    d. Tujuan
    e. Metodologi
    f. Sistematika Penulisan
2. Tinjauan Pustaka
3. Desain dan Implementasi
4. Pengujian dan Evaluasi
5. Kesimpulan dan Saran
6. Daftar Pustaka

11. JADWAL KEGIATAN

Jadwal kegiatan pada Tugas Akhir ini akan dijelaskan pada Tabel 1.

Tabel 1. Rencana Jadwal Kegiatan

Tahapan	2014																			
	Februari				Maret				April				Mei					Juni		
Penyusunan Proposal																				
Studi Literatur																				
Perancangan sistem																				
Implementasi																				
Pengujian dan evaluasi																				
Penyusunan buku																				

12. DAFTAR PUSTAKA

[1] N. Nguyen and R. Caruana. (2007, Oct.) Consensus Clusterings. [Online]. http://www.cs.cornell.edu/

[2] A. Lourenço, et al., "Consensus Clustering Using Partial Evidence Accumulation," *Pattern Recognition and Image Analysis*, pp. 69-78, 2013.

[3] A. L. Fred, Member, IEEE, a. A. K. Jain, and Fellow, "Combining Multiple Clusterings Using Evidence Accumulation," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 27, Jun. 2005.

[4] A. K. Jain and A. L. N. Fred, "Data Clustering Using Evidence Accumulation," *IEEE Pattern Recognition*, vol. 4, pp. 276-280, 2002.