



**PROGRAM STUDI
TEKNIK INFORMATIKA – S1
FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO**

MATA KULIAH
DATA MINING



Technology vector created by sentavio - www.freepik.com

DATA MINING

“Klastering”

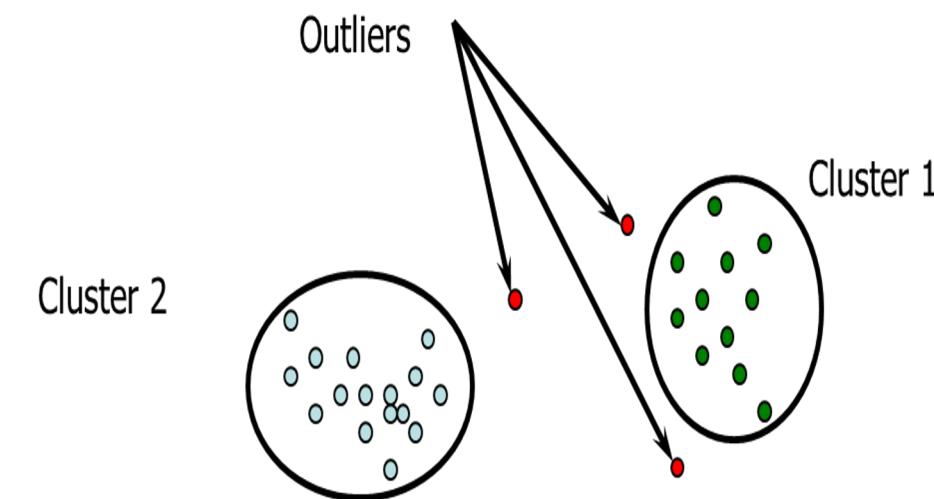
TIM PENGAMPU DOSEN DATA MINING

Introduction

- Clustering berkaitan dengan pengelompokan bersama objek yang mirip satu sama lain dan berbeda dengan objek milik cluster lain (Bramer, 2020).
- Analisis cluster adalah metode data mining deskriptif yang paling terkenal. Dengan adanya matriks data yang terdiri dari n pengamatan (baris) dan variabel p (kolom), tujuan dari analisis cluster adalah untuk mengelompokkan pengamatan menjadi kelompok-kelompok yang secara internal homogen (kohesi internal) dan heterogen dari kelompok ke kelompok (pemisahan eksternal). (Guidici, 2009).
- Yang kami maksud dengan pengelompokan adalah metode untuk membagi sekumpulan data (catatan / tuple / vektor / instance / objek / sampel) menjadi beberapa kelompok (cluster), berdasarkan kesamaan tertentu yang telah ditentukan sebelumnya. (Goronescu, 2011).

Introduction [2]

- Pengelompokan mengacu pada pengelompokan rekaman, observasi, atau kasus ke dalam kelas objek serupa. Cluster adalah kumpulan record yang serupa satu sama lain, dan berbeda dengan record di cluster lain.
- Algoritma pengelompokan berusaha untuk menyegmentasikan seluruh kumpulan data menjadi subkelompok atau cluster yang relatif homogen, di mana kesamaan catatan dalam cluster dimaksimalkan dan kesamaan dengan catatan di luar cluster diminimalkan (Larose, 2005).



Algoritma Klastering

- Klastering adalah pengelompokkan data, hasil observasi dan kasus ke dalam class yang mirip.
- Suatu klaster (cluster) adalah koleksi data yang mirip antara satu dengan yang lain, dan memiliki perbedaan bila dibandingkan dengan data dari klaster lain.
- Perbedaan utama algoritma klastering dengan klasifikasi adalah klastering tidak memiliki target/class/label, jadi termasuk *unsupervised learning*.
- Klastering sering digunakan sebagai tahap awal dalam proses data mining, dengan hasil klaster yang terbentuk akan menjadi input dari algoritma berikutnya yang digunakan.

Jenis – Jenis Klastering

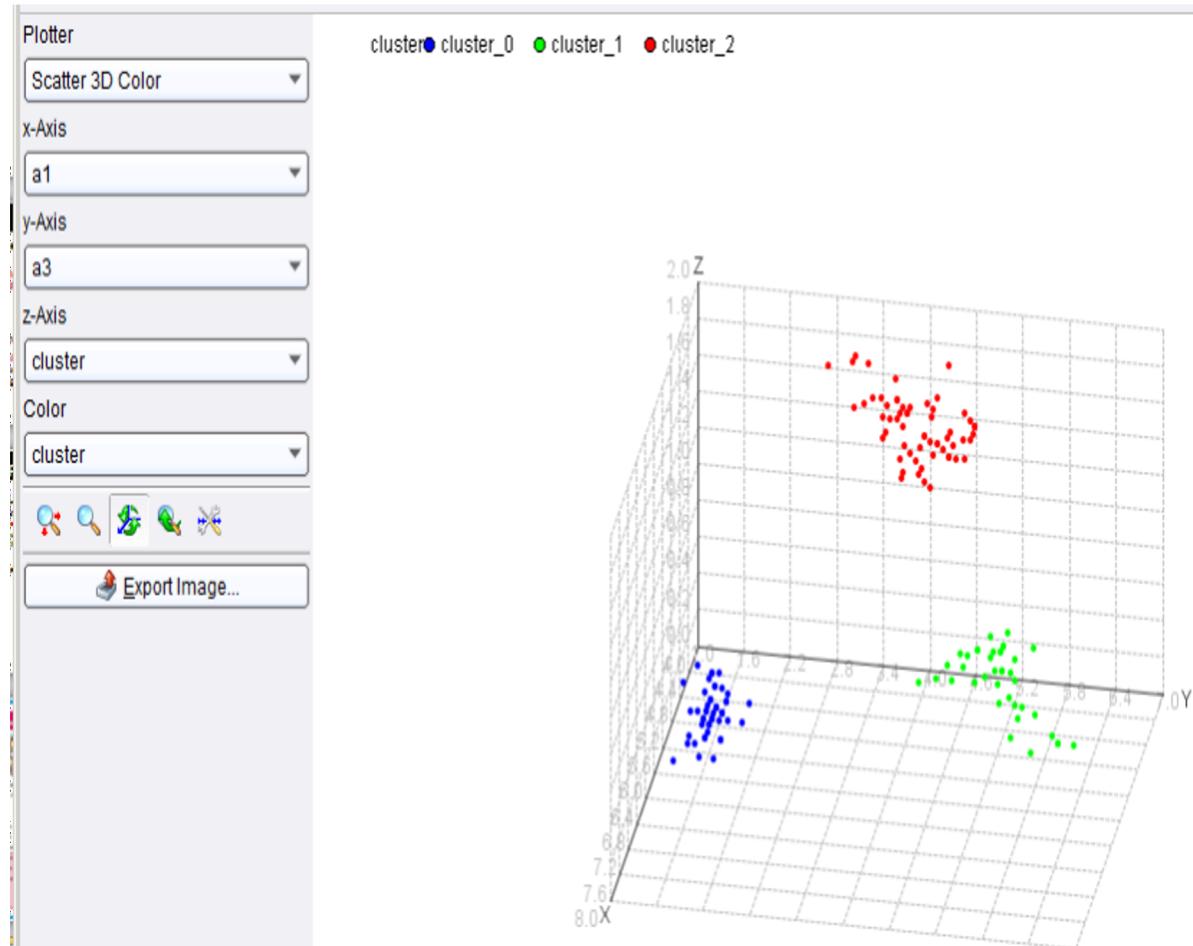
William membagi algoritma *clustering* ke dalam kelompok besar seperti berikut:

1. *Partitioning algorithms*: algoritma dalam kelompok ini membentuk bermacam partisi dan kemudian mengevaluasinya dengan berdasarkan beberapa kriteria.
2. *Hierarchy algorithms*: pembentukan dekomposisi hirarki dari sekumpulan data menggunakan beberapa kriteria.
3. *Density based*: pembentukan cluster berdasarkan pada koneksi dan fungsi densitas.
4. *Grid-based*: pembentukan cluster berdasarkan pada struktur multiple level granularity
5. *Model-based*: sebuah model dianggap sebagai hipotesa untuk masing-masing cluster dan model yang baik dipilih diantara model hipotesa tersebut.

Contoh Klastering Bunga Iris

ExampleSet (150 examples, 2 special attributes, 4 regular attributes)						
Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_3	Iris-setosa	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	5	3.600	1.400	0.200
6	id_6	Iris-setosa	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	5	3.400	1.500	0.200
9	id_9	Iris-setosa	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	4.800	3	1.400	0.100
14	id_14	Iris-setosa	4.300	3	1.100	0.100
15	id_15	Iris-setosa	5.800	4	1.200	0.200
16	id_16	Iris-setosa	5.700	4.400	1.500	0.400
17	id_17	Iris-setosa	5.400	3.900	1.300	0.400
18	id_18	Iris-setosa	5.100	3.500	1.400	0.300
19	id_19	Iris-setosa	5.700	3.800	1.700	0.300
20	id_20	Iris-setosa	5.100	3.800	1.500	0.300
21	id_21	Iris-setosa	5.400	3.400	1.700	0.200
22	id_22	Iris-setosa	5.100	3.700	1.500	0.400
23	id_23	Iris-setosa	4.600	3.600	1	0.200
24	id_24	Iris-setosa	5.100	3.300	1.700	0.500

Contoh: Klastering Bunga Iris (Plot)



Contoh: Klastering Bunga Iris (Table)

ExampleSet (150 examples, 3 special attributes, 4 regular attributes)							
Row No.	id	label	cluster	a1	a2	a3	a4
1	id_1	Iris-setosa	cluster_0	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	cluster_0	4.900	3	1.400	0.200
3	id_3	Iris-setosa	cluster_0	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	cluster_0	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	cluster_0	5	3.600	1.400	0.200
6	id_6	Iris-setosa	cluster_0	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	cluster_0	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	cluster_0	5	3.400	1.500	0.200
9	id_9	Iris-setosa	cluster_0	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	cluster_0	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	cluster_0	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	cluster_0	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	cluster_0	4.800	3	1.400	0.100
14	id_14	Iris-setosa	cluster_0	4.300	3	1.100	0.100
15	id_15	Iris-setosa	cluster_0	5.800	4	1.200	0.200
16	id_16	Iris-setosa	cluster_0	5.700	4.400	1.500	0.400
17	id_17	Iris-setosa	cluster_0	5.400	3.900	1.300	0.400
18	id_18	Iris-setosa	cluster_0	5.100	3.500	1.400	0.300
19	id_19	Iris-setosa	cluster_0	5.700	3.800	1.700	0.300
20	id_20	Iris-setosa	cluster_0	5.100	3.800	1.500	0.300

Cluster Model

Cluster 0: 50 items

Cluster 1: 39 items

Cluster 2: 61 items

Total number of items: 150

Latihan Soal (Kuis)

- Carilah 3 paper tentang perkembangan aplikasi Data Mining menggunakan Klastering minimal 5 tahun terakhir, kemudian review paper tersebut. Korelasikan dengan hal yang yang paling memungkinkan ada di sekitar anda.

Referensi

1. Kusrini, Taufiq Emha, Algoritma Data Mining, *Penerbit Andi*, 2009.
2. Ian H. Witten, Frank Eibe, Mark A. Hall, Data mining: Practical Machine Learning Tools and Techniques 4th Edition, *Elsevier*, 2017.
3. Budi Santosa, Ardian Umam, Data Mining dan Big Data Analytics, Penebar Media Pustaka, 2018.
4. Yaya Heryadi, Teguh Wahyono, Machine Learning: Konsep dan Implementasi, Penerbit Gava Media, 2020.
5. Sumber gambar: www.freepik.com.

TUGAS AKHIR

- Tema / Judul Eksperimen (Point of Interest)
- Literature (Paper, Jurnal, Artikel, Buku)
- Dataset
- Tahapan Eksperimen
- Git dan Repository
- Publish / Deployment

Mulai report progress 2 minggu kedepan (Report pada kulino dan Diskusi kelas)



THANKS

ANY QUESTIONS?

