**MAKALAH**

**KLASTERING TWEET MENGGUNAKAN BAHASA (R)**



**Disusun Oleh :**

**Agus Mudiyanto (14.01.53.0040)**

**Mukhamad Iqbal Muzadi (14.01.53.0036)**

**Vikko Dwi Prakoso (14.01.53.0036)**

**UNIVERSITAS STIKUBANK SEMARANG (UNISBANK)**

**FALKUTAS TEKNOLOGI INFORMASI**

**SISTEM INFORMASI**

**2017**

**BAB I**

**PENDAHULUAN**

1. **Latar belakang**

Clustering adalah metode penganalisaan data, yang sering dimasukkan sebagai salah satu metode *Data Mining*, yang tujuannya adalah untuk mengelompokkan data dengan karakteristik yang sama ke suatu ‘wilayah’ yang sama dan data dengan karakteristik yang berbeda ke ‘wilayah’ yang lain. Ada beberapa pendekatan yang digunakan dalam mengembangkan metode clustering. Dua pendekatan utama adalah clustering dengan pendekatan partisi dan clustering dengan pendekatan hirarki. Clustering dengan pendekatan partisi atau sering disebut dengan *partition-based clustering* mengelompokkan data dengan memilah-milah data yang dianalisa ke dalam cluster-cluster yang ada. Clustering dengan pendekatan hirarki atau sering disebut dengan *hierarchical clustering*mengelompokkan data dengan membuat suatu hirarki berupa dendogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hirarki yang berjauhan. Di samping kedua pendekatan tersebut, ada juga clustering dengan pendekatan *automatic mapping*(Self-Organising Map/SOM).

Twitter merupakan situs jejaring sosial yang keberadaannya masih diminiati oleh masyarkat sampai saat ini. Twitter adalah jejaring sosial berupa blog ukuran kecil yang didirikan oleh Jack Dorsey pada bulan Maret 2006. Melalui Twitter pengguna dapat mengirim dan membaca pesan, berbagi informasi, menjalin relasi bisnis, menuangkan isi hati dan pikiran dalam bentuk tulisan (sering disebut tweet), dengan kapasitas kata yang bisa diunggah dan ditampilkan pada timeline penggunatwitter mencapai 140 karakter. Sama halnya dengan situs jejaring sosial lain dalam Twitter disediakan suatu mesin pencarian (search engine) yang berguna untuk mempermudah pengguna dalam menemukan informasi menggunakan kata kunci. Melalui search engine pengguna dapat menemukan lebih banyak informasi yang dibutuhkan terkait topik yang ingin dicari, yaitu lebih dari satu akun yang ada di twitter.Twitter sebagai hasil dari perkembangan teknologi informasi memungkinkan setiap waktu untuk menghasilkan kumpulan data yang banyak, dimana setiap detiknya pada saat kehidupan normal rata-rata jumlah tweet yang ada dalam twitter adalah 600 tweet. Dengan adanya kumpulan data yang terus meningkat setiap waktunya yaitu berupa data tweet perlu dilakukan suatu penangan menggunakan metode khusus untuk mengalisis data pada twitter sehingga menghasilkan suatu informasi yang bermanfaat dan mengurangi kondisi yang biasa disebut “rich of data but poor of information”.

* 1. **Rumusan Masalah**

Berdasarkan latar belakang di atas maka rumusan masalahnya adalah bagaimana penjelasan mengenai Klustering clustering dengan pendekatan partisi, clustering dengan pendekatan hirarki, clustering dengan pendekatan automatic mapping pada tweet dan implementasinya menggunakan bahasa R.

* 1. **Tujuan Penulisan**

Adapun tujuan dari penulisan ini adalah untuk mengklasifikasikan Klustering pada sebuah twitter menggunakan bahasa R.

**BAB II**

**PEMBAHASAN**

* 1. [**Clustering**](https://yudiagusta.wordpress.com/clustering/)

Clustering adalah metode penganalisaan data, yang sering dimasukkan sebagai salah satu metode *Data Mining*, yang tujuannya adalah untuk mengelompokkan data dengan karakteristik yang sama ke suatu ‘wilayah’ yang sama dan data dengan karakteristik yang berbeda ke ‘wilayah’ yang lain.

Ada beberapa pendekatan yang digunakan dalam mengembangkan metode clustering. Dua pendekatan utama adalah clustering dengan pendekatan partisi dan clustering dengan pendekatan hirarki. Clustering dengan pendekatan partisi atau sering disebut dengan *partition-based clustering* mengelompokkan data dengan memilah-milah data yang dianalisa ke dalam cluster-cluster yang ada. Clustering dengan pendekatan hirarki atau sering disebut dengan *hierarchical clustering* mengelompokkan data dengan membuat suatu hirarki berupa dendogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hirarki yang berjauhan. Di samping kedua pendekatan tersebut, ada juga clustering dengan pendekatan *automatic mapping* (Self-Organising Map/SOM)

**2.1.1 Clustering Dengan Pendekatan Partisi**

*K-Means*

Salah satu metode yang banyak digunakan dalam melakukan clustering dengan partisi ini adalah metode k-means. Secara umum metode k-means ini melakukan proses pengelompokan dengan prosedur sebagai berikut:

* Tentukan jumlah cluster
* Alokasikan data secara random ke cluster yang ada
* Hitung rata-rata setiap cluster dari data yang tergabung di dalamnya
* Alokasikan kembali semua data ke cluster terdekat
* Ulang proses nomor 3, sampai tidak ada perubahan atau perubahan yang terjadi masih sudah di bawah treshold

Prosedur dasar ini bisa berubah mengikuti pendekatan pengalokasian data yang diterapkan, apakah *crisp* atau *fuzzy*. Setelah meneliti clustering dari sudut yang lain, saya menemukan bahwa *k-means clustering* mempunyai beberapa kelemahan.  
Hal-hal terkait dengan metode k-means saya rangkum dalam tulisan saya yang dapat di-download di sini ([k-means – penerapan, permasalahan dan metode terkait](https://yudiagusta.files.wordpress.com/2008/03/k-means.pdf)).

Penjelasan lengkap tentang k-means dapat juga dilihat pada [Yudi Agusta’s K-Means Page](https://yudiagusta.wordpress.com/k-means/).

*Mixture Modelling (Mixture Modeling)*

Mixture modelling (mixture modeling) merupakan metode pengelompokan data yang mirip dengan k-means dengan kelebihan penggunaan distribusi statistik dalam mendefinisikan setiap cluster yang ditemukan. Dibandingkan dengan k-means yang hanya menggunakan cluster center, penggunaan distribusi statistik ini mengijinkan kita untuk:

* Memodel data yang kita miliki dengan setting karakteristik yang berbeda-beda
* Jumlah cluster yang sesuai dengan keadaan data bisa ditemukan seiring dengan proses pemodelan karakteristik dari masing-masing cluster
* Hasil pemodelan clustering yang dilaksanakan bisa diuji tingkat keakuratannya

Distribusi statistik yang digunakan bisa bermacam-macam mulai dari yang digunakan untuk data categorical sampai yang continuous, termasuk di antaranya distribusi binomial, multinomial, normal dan lain-lain. Beberapa distribusi yang bersifat tidak normal seperti distribusi Poisson, von-Mises, Gamma dan Student t, juga diterapkan untuk bisa mengakomodasi berbagai keadaan data yang ada di lapangan. Beberapa pendekatan multivariate juga banyak diterapkan untuk memperhitungkan tingkat keterkaitan antara variabel data yang satu dengan yang lainnya.

K-means merupakan salah satu algoritma clustering .Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan supervised learning yang menerima masukan berupa vektor (­x­1 , y1) , (­x­2 , y2) , …, (­x­i , yi), di mana xi merupakan data dari suatu data pelatihan dan yi merupakan label kelas untuk xi.

Pada algoritma pembelajaran ini, komputer mengelompokkan sendiri data-data yang menjadi masukannya tanpa mengetahui terlebih dulu target kelasnya. Pembelajaran ini termasuk dalam unsupervised learning. Masukan yang diterima adalah data atau objek dan k buah kelompok (cluster) yang diinginkan. Algoritma ini akan mengelompokkan data atau objek ke dalam k buah kelompok tersebut. Pada setiap cluster terdapat titik pusat (centroid) yang merepresentasikan cluster tersebut.

Algoritma untuk melakukan K-Means clustering adalah sebagai berikut:

* Pilih K buah titik centroid secara acak
* Kelompokkan data sehingga terbentuk K buah cluster dengan titik centroid dari
* setiap cluster merupakan titik centroid yang telah dipilih sebelumnya

Perbaharui nilai titik centroid

* Ulangi langkah 2 dan 3 sampai nilai dari titik centroid tidak lagi berubah

Proses pengelompokkan data ke dalam suatu cluster dapat dilakukan dengan cara menghitung jarak terdekat dari suatu data ke sebuah titik centroid. Perhitungan jarak Minkowski dapat digunakan untuk menghitung jarak antar 2 buah data. Rumus untuk menghitung jarak tersebut adalah:



Di mana:

g = 1, untuk menghitung jarak Manhattan

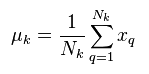
g = 2, untuk menghitung jarak Euclidean

g = ∞, untuk menghitung jarak Chebychev

xi , xj adalah dua buah data yang akan dihitung jaraknya

p = dimensi dari sebuah data

Pembaharuan suatu titik centroid dapat dilakukan dengan rumus berikut:



Di mana:

µk = titik centroid dari cluster ke-K

Nk = banyaknya data pada cluster ke-K

xq = data ke-q pada cluster ke-K

**Kekurangan dan Kelebihan K-Means**

Ada beberapa kelebihan pada algoritma k-means, yaitu:

1. Mudah untuk diimplementasikan dan dijalankan.
2. Waktu yang dibutuhkan untuk menjalankan pembelajaran ini relatif cepat.
3. Mudah untuk diadaptasi.
4. Umum digunakan.

Algoritma k-means memiliki beberapa kelebihan, namun ada kekurangannya juga. Kekurangan dari algoritma tersebut yaitu :

1. Sebelum algoritma dijalankan, k buah titik diinisialisasi secara random sehingga pengelompokkan data yang dihasilkan dapat berbeda-beda. Jika nilai random untuk inisialisasi kurang baik, maka pengelompokkan yang dihasilkan pun menjadi kurang optimal.
2. Dapat terjebak dalam masalah yang disebut curse of dimensionality. Hal ini dapat terjadi jika data pelatihan memiliki dimensi yang sangat tinggi (Contoh jika data pelatihan terdiri dari 2 atribut maka dimensinya adalah 2 dimensi. Namun jika ada 20 atribut, maka akan ada 20 dimensi). Salah satu cara kerja algoritma ini adalah mencari jarak terdekat antara k buah titik dengan titik lainnya. Jika mencari jarak antar titik pada 2 dimensi, masih mudah dilakukan. Namun bagaimana mencari jarak antar titik jika terdapat 20 dimensi. Hal ini akan menjadi sulit.

Jika hanya terdapat beberapa titik sampel data, maka cukup mudah untuk menghitung dan mencari titik terdekat dengan k titik yang diinisialisasi secara random. Namun jika terdapat banyak sekali titik data (misalnya satu milyar buah data), maka perhitungan dan pencarian titik terdekat akan membutuhkan waktu yang lama. Proses tersebut dapat dipercepat, namun dibutuhkan struktur data yang lebih rumit seperti kD-Tree atau hashing.

**2.1.2 Clustering dengan Pendekatan Hirarki**

Clustering dengan pendekatan hirarki mengelompokkan data yang mirip dalam hirarki yang sama dan yang tidak mirip di hirarki yang agak jauh. Ada dua metode yang sering diterapkan yaitu *agglomerative hieararchical clustering* dan *divisive hierarchical clustering*. Agglomerative melakukan proses clustering dari *N* cluster menjadi satu kesatuan cluster, dimana *N* adalah jumlah data, sedangkan divisive melakukan proses clustering yang sebaliknya yaitu dari satu cluster menjadi *N* cluster.

Beberapa metode *hierarchical clustering* yang sering digunakan dibedakan menurut cara mereka untuk menghitung tingkat kemiripan. Ada yang menggunakan *Single Linkage*, *Complete Linkage*, *Average Linkage*, *Average Group Linkage* dan lain-lainnya. Seperti juga halnya dengan *partition-based clustering*, kita juga bisa memilih jenis jarak yang digunakan untuk menghitung tingkat kemiripan antar data.

Salah satu cara untuk mempermudah pengembangan dendogram untuk hierarchical clustering ini adalah dengan membuat similarity matrix yang memuat tingkat kemiripan antar data yang dikelompokkan. Tingkat kemiripan bisa dihitung dengan berbagai macam cara seperti dengan Euclidean Distance Space. Berangkat dari similarity matrix ini, kita bisa memilih lingkage jenis mana yang akan digunakan untuk mengelompokkan data yang dianalisa.

**2.1.3 Clustering Dengan Pendekatan Automatic Mapping**

Self-Organising Map (SOM) merupakan suatu tipe Artificial Neural Networks yang di-training secara unsupervised. SOM menghasilkan map yang terdiri dari output dalam dimensi yang rendah (2 atau 3 dimensi). Map ini berusaha mencari property dari input data. Komposisi input dan output dalam SOM mirip dengan komposisi dari proses feature scaling (multidimensional scaling).

Walaupun proses learning yang dilakukan mirip dengan Artificial Neural Networks, tetapi proses untuk meng-assign input data ke map, lebih mirip dengan K-Means dan kNN Algorithm. Adapun prosedur yang ditempuh dalam melakukan clustering dengan SOM adalah sebagai berikut:

* Tentukan weight dari input data secara random
* Pilih salah satu input data
* Hitung tingkat kesamaan (dengan Eucledian) antara input data dan weight dari input data tersebut dan pilih input data yang memiliki kesamaan dengan weight yang ada (data ini disebut dengan Best Matching Unit (BMU))
* Perbaharui weight dari input data dengan mendekatkan weight tersebut ke BMU dengan rumus:

*Wv(t+1) = Wv(t) + Theta(v, t) x Alpha(t) x (D(t) – Wv(t))*

Dimana:

* + *Wv(t)*: Weight pada saat ke-t
  + *Theta (v, t)*: Fungsi neighbourhood yang tergantung pada Lattice distance antara BMU dengan neuron v. Umumnya bernilai 1 untuk neuron yang cukup dekat dengan BMU, dan 0 untuk yang sebaliknya. Penggunaan fungsi Gaussian juga memungkinkan.
  + *Alpha (t)*: Learning Coefficient yang berkurang secara monotonic
  + *D(t)*: Input data
* Tambah nilai t, sampai t < *Lambda*, dimana *Lambda* adalah jumlah iterasi

**Variasi Metode Clustering**

* [*Quality Threshold Clustering Method*](https://yudiagusta.wordpress.com/2008/06/12/quality-threshold-clustering/)
* [*Locality Sensitive Hashing*](https://yudiagusta.wordpress.com/2009/09/07/locality-sensitive-hashing/)
* [Algoritma Rock](https://yudiagusta.wordpress.com/2009/10/05/algoritma-rock/)
* Hierarchical Frequent Term-Base Clustering
* Suffix Tree Clustering
* [Single Pass Clustering](https://yudiagusta.wordpress.com/2010/07/19/single-pass-clustering/)
* Neighborhood Clustering
* [Sequence Clustering](https://yudiagusta.wordpress.com/2008/06/24/sequence-clustering/)
* Spectral Clustering
* Clustering on Frequent Tree
* [Latent Class Cluster Analysis](https://yudiagusta.wordpress.com/2009/07/29/latent-class-cluster-analysis/) a.k.a. Latent Profile Analysis a.k.a. [Mixture Model](https://yudiagusta.wordpress.com/mixture-modelling/) for Continuous Variabel
* [Latent Class Analysis](https://yudiagusta.wordpress.com/2009/11/02/latent-class-analysis/) a.k.a. [Mixture Model](https://yudiagusta.wordpress.com/mixture-modelling/) for Categorical Variable

**Hal-hal Terkait Dengan Clustering**

* Analisa Faktor
* Singular Value Decomposition
* Eigen Value and Eigen Vector
* [Similarity Measure](https://yudiagusta.wordpress.com/2008/05/13/similarity-measure/)
* Feature Discretisation
* [Feature Selection](https://yudiagusta.wordpress.com/2009/01/13/feature-selection/)
* Feature Scaling
* Indexing Method For Searching

**BAB III**

**PENUTUP**

**Kesimpulan**

Klustering adalah mengelompokkan variabel ke dalam kelompok berdasarkan kesamaan tertentu.Sedangkan R adalah suatu perangkat lunak yang digunakan untuk manipulasi data, perhitungan, simulasi, penayangan grafik, dan sekaligus sebagai bahasa pemograman yang bersifat interpreter.R adalah bahasa pemograman yang bersifat object oriented, sehingga seluruh variable, data, fungsi, dan lain sebagainya disimpan dalam memori komputer sebagai objek.Program merupakan serangkaian instruksi yang memberitahu komputer sesuatu yang harus dilakukan (Zelle, 2002).Pemrograman dengan bahasa R dapat mengklasifikasikan Klustering K-means,PAM(Partitioning AroundMedoids),Hirarki Agglomerative pada sebuah twitter menggunakan bahasa R.