Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster)
Volume 02 , No. 2 (2013), hal 133 - 136

PENGKLASIFIKASIAN KARAKTERISTIK DENGAN METODE *K-MEANS CLUSTER ANALYSIS*

Ediyanto, Muhlasah Novitasari Mara, Neva Satyahadewi

INTISARI

Pengelompokan objek (objek clustering) adalah salah satu proses dari objek mining yang bertujuan untuk mempartisi objek yang ada kedalam satu atau lebih cluster objek berdasarkan karakteristiknya. Objek dengan karakteristik yang sama dikelompokkan dalam satu cluster dan objek dengan karakteristik berbeda dikelompokkan kedalam cluster yang lain. Algoritma K-Means Cluster Analysis termasuk dalam kelompok metode cluster analysis non hirarki, dimana jumlah kelompok yang akan dibentuk sudah terlebih dahulu diketahui atau ditetapkan jumlahnya. Algoritma K-Means Cluster Analysis menggunakan metode perhitungan jarak (distance) untuk mengukur tingkat kedekatan antara objek dengan titik tengah (centroid). Algoritma K-Means tidak terpengaruh terhadap urutan objek yang digunakan, hal ini dibuktikan ketika penulis mencoba menentukan secara acak titik awal pusat cluster dari salah satu objek pada permulaan perhitungan. Jumlah keanggotaan cluster yang dihasilkan berjumlah sama ketika menggunakan objek yang lain sebagai titik awal pusat cluster tersebut. Namun, hal ini hanya berpengaruh pada jumlah iterasi yang dilakukan. Algoritma K-Means Cluster Analysis pada dasarnya dapat diterapkan pada permasalahan dalam memahami perilaku konsumen, mengidentifikasi peluang produk baru dipasaran dan algoritma K-Means ini juga dapat digunakan untuk meringkas objek dari jumlah besar sehingga lebih memudahkan untuk mendiskripsikan sifat-sifat atau karakteristik dari masing-masing kelompok.

Kata Kunci: Clustering, Cluster Analysis, Euclidian distance, K-Means.

PENDAHULUAN

Analisis *Cluster* merupakan teknik multivariat yang mempunyai tujuan utama untuk mengelompokan objek-objek berdasarkan karakteristik yang dimilikinya. Analisis *Cluster* mengklasifikasi objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam *cluster* yang sama. *Cluster-cluster* yang terbentuk memiliki homogenitas internal yang tinggi dan heterogenitas eksternal yang tinggi. Berbeda dengan teknik multivariat lainnya, analisis ini tidak mengestimasi set variabel secara empiris sebaliknya menggunakan set variabel yang ditentukan oleh peneliti itu sendiri. Fokus dari Analisis *Cluster* adalah membandingkan objek berdasarkan set variabel, hal inilah yang menyebabkan para ahli mendefinisikan set variabel sebagai tahap kritis dalam analisis *cluster*. Set variabel *cluster* adalah suatu set variabel yang merpresentasikan karakteristik yang dipakai objek-objek. Solusi Analisis *Cluster* bersifat tidak unik, anggota *cluster* untuk tiap penyelesaian/solusi tergantung pada beberapa elemen prosedur dan beberapa solusi yang berbeda dapat diperoleh dengan mengubah satu elemen atau lebih. Solusi *cluster* secara keseluruhan bergantung pada variabel-variabel yang digunakan sebagai dasar untuk menilai kesamaan. Penambahan atau pengurangan variabel-variabel yang relevan dapat mempengaruhi substansi hasil analisis *cluster*. Pada tulisan ini penulis menggunakan metode *K-Means Cluster Analysis* sebagai solusi untuk pengklasifikasian karakteristik dari objek. Alasan penggunaan algoritma *K-Means* diantaranya ialah karena algoritma ini memiliki ketelitian yang cukup tinggi terhadap ukuran objek, sehingga algoritma ini relatif lebih terukur dan efisien untuk pengolahan objek dalam jumlah besar. Selain itu algoritma *K-Means* ini tidak terpengaruh terhadap urutan objek [1].

Permasalahan yang dikaji dalam tulisan ini adalah bagaimana penggunaan metode *K-Means Cluster Analysis* dalam pengklasifikasian karakteristik suatu objek, tujuan yang ingin penulis capai

adalah mengkaji metode *K-Means Cluster Analysis* dalam pengklasifikasian karakteristik berdasarkan set variabel yang dibentuk. Metode ukuran jarak yang digunakan dalam menghitung jarak objek terhadap *centroid* yaitu persamaan jarak *Euclidian*. Pada algoritma *K-Means Cluster Analysis* terdapat beberapa langkah yang harus dilakukan yaitu sebagai berikut:

1. Tentukan jumlah *cluster*.
2. Alokasikan objek ke dalam *cluster* secara random.
3. Hitung *centroid* sampel yang ada di masing-masing *cluster*.
4. Alokasikan masing-masing objek ke *centroid* terdekat.
5. Kembali ke langkah 3 apabila masih ada objek yang berpindah *cluster* atau masih ada perubahan nilai *centroid*, ada yang di atas nilai *threshold* yang ditentukan atau apabila perubahan nilai pada *objective function* yang digunakan di atas nilai *threshold* yang ditentukan [2].

Sebagai bahan bacaan dalam tulisan ini penulis juga membandingkan beberapa karya ilmiah yang mengupas materi yang sama tentang Analisis *Cluster*, Rismawan dan Kusumadewi dalam penelitiannya mencoba membangun suatu sistem untuk mengelompokan objek yang ada berdasarkan status gizi dan ukuran rangka dari objek yang diambil dengan memasukkan parameter kondisi fisik dari objek orang tersebut. Pengelompokan objek dilakukan dengan menggunakan metode *K-Means Cluster* yaitu mengelompokan *n* buah objek ke dalam *k* kelas berdasarkan jaraknya dengan pusat kelas. Hasil dari penelitian ini terhadap 20 objek sampel diperoleh 3 kelompok mahasiswa berdasarkan nilai BMI (*Body Mass Index*) dan ukuran rangka, yaitu BMI normal dengan kerangka besar, BMI obesitas sedang dengan kerangka sedang, dan BMI obesitas berat dengan kerangka kecil [3].

Dalam penelitian kuantitatif, populasi dan sampel merupakan sumber utama untuk memperoleh objek yang dibutuhkan pada suatu penelitian survei (*survey research*). Untuk mencapai keakuratan dan validitas objek yang dihasilkan, maka kejelasan dan ketepatan dalam pengambilan sampel sangat diprioritaskan baik dari segi ukuran atau besarnya sampel maupun karakteristik yang dimilikinya. Populasi atau sering juga disebut *universe* adalah keseluruhan atau totalitas objek yang diteliti yang ciri-cirinya akan diduga atau ditaksir [4]. Ciri-ciri populasi disebut parameter. Oleh karena itu, populasi juga sering diartikan sebagai kumpulan objek penelitian dari mana objek akan dijaring atau dikumpulkan. Sampel adalah bagian dari populasi yang diambil untuk keperluan analisis [5]. Apabila populasi berukuran besar, peneliti biasanya kesulitan untuk mempelajari semua karakteristik yang ada pada populasi, hal ini mungkin dikarenakan adanya keterbatasan dana, tenaga dan waktu, maka peneliti dapat menggunakan sampel yang diambil dari populasi tersebut. Apa yang dipelajari dari sampel itu, kesimpulannya akan diberlakukan untuk populasi. Oleh karena itu sampel yang diambil dari populasi harus betul-betul dapat mewakili (*representative*). Sampel *representative* adalah sampel yang memiliki ciri karakteristik yang sama atau relatif sama dengan ciri karakteristik populasinya. Tingkat kerepresentatifan sampel yang diambil dari populasi tertentu sangat tergantung pada jenis sampel yang digunakan, ukuran sampel yang diambil, dan cara pengambilannya. Cara atau prosedur yang digunakan untuk mengambil sampel dari populasi tertentu disebut teknik *sampling*.

K-MEANS CLUSTER ANALYSIS

Cluster Analysis merupakan salah satu metode *objek mining* yang bersifat tanpa latihan (*unsupervised analysis*), sedangkan *K-Means Cluster Analysis* merupakan salah satu metode *cluster analysis* non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih *cluster* atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu *cluster* yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan kedalam *cluster* yang lain. Tujuan *pengelompokan* adalah untuk meminimalkan *objective function* yang di set dalam proses *clustering*, yang pada dasarnya berusaha untuk meminimalkan variasi dalam satu *cluster* dan memaksimalkan variasi antar *cluster*.

Metode *cluster* ini meliputi *sequential threshold, pararel threshold dan optimizing threshold. Sequential threshold* melakukan pengelompokan dengan terlebih dahulu memilih satu objek dasar yang akan dijadikan nilai awal *cluster*, kemudian semua *cluster* yang ada dalam jarak terdekat dengan *cluster* ini akan bergabung, lalu dipilih *cluster* kedua dan semua objek yang mempunyai kemiripan dengan *cluster* ini akan digabungkan, demikian seterusnya sehingga terbentuk beberapa *cluster* dengan keseluruhan objek yang terdapat didalamnya.

Jika diberikan sekumpulan objek $X = (x_1, x_2, \dots, x_n)$ maka algoritma *K-Means Cluster Analysis* akan mempartisi *X* dalam *k* buah *cluster*, setiap *cluster* memiliki *centroid* dari objek-objek dalam *cluster* tersebut. Pada tahap awal algoritma *K-Means Cluster Analysis* dipilih secara acak *k* buah objek sebagai *centroid*, kemudian jarak antara objek dengan *centroid* dihitung dengan menggunakan jarak *euclidian,* objek ditempatkan dalam *cluster* yang terdekat dihitung dari titik tengah *cluster*. *Centroid* baru ditetapkan jika semua objek sudah ditempatkan dalam *cluster* terdekat. Proses penentuan *centroid* dan penempatan objek dalam *cluster* diulangi sampai nilai *centroid* konvergen (*centroid* dari semua *cluster* tidak berubah lagi). Secara umum metode *K-Means Cluster Analysis* menggunakan algoritma sebagai berikut [6]:

1. Tentukan *k* sebagai jumlah *cluster* yang di bentuk.
   Untuk menentukan banyaknya *cluster k* dilakukan dengan beberapa pertimbangan seperti pertimbangan teoritis dan konseptual yang mungkin diusulkan untuk menentukan berapa banyak *cluster*.
2. Bangkitkan *k Centroid* (titik pusat *cluster*) awal secara *random.*
   Penentuan *centroid* awal dilakukan secara *random*/acak dari objek-objek yang tersedia sebanyak *k cluster*, kemudian untuk menghitung *centroid cluster* ke-*i* berikutnya, digunakan rumus sebagai berikut :

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \quad ; i = 1,2,3,\dots n \qquad (1)$$

   dimana; v : *centroid* pada *cluster*
   x_i : objek ke-*i*
   n : banyaknya objek/jumlah objek yang menjadi anggota *cluster*
3. Hitung jarak setiap objek ke masing-masing *centroid* dari masing-masing *cluster*.
   Untuk menghitung jarak antara objek dengan *centroid* penulis menggunakan *Euclidian Distance.*

$$d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad ; i = 1,2,3,\dots, n \qquad (2)$$

   dimana x_i : objek *x* ke-i
   y_i : daya *y* ke-i
   n : banyaknya objek
4. Alokasikan masing-masing objek ke dalam *centroid* yang paling terdekat.
   Untuk melakukan pengalokasian objek kedalam masing-masing *cluster* pada saat iterasi secara umum dapat dilakukan dengan dua cara yaitu dengan *hard k-means*, dimana secara tegas setiap objek dinyatakan sebagai anggota *cluster* dengan mengukur jarak kedekatan sifatnya terhadap titik pusat *cluster* tersebut, cara lain dapat dilakukan dengan *fuzzy C-Means.*
5. Lakukan iterasi, kemudian tentukan posisi *centroid* baru dengan menggunakan persamaan (1).
6. Ulangi langkah 3 jika posisi *centroid* baru tidak sama.
   Pengecekan *konvergensi* dilakukan dengan membandingkan matriks *group assignment* pada iterasi sebelumnya dengan matrik *group assignment* pada iterasi yang sedang berjalan. Jika hasilnya sama maka *algoritma k-means cluster analysis* sudah *konvergen*, tetapi jika berbeda maka belum *konvergen* sehingga perlu dilakukan iterasi berikutnya.

Pada penerapan metode *K-Means Cluster Analysis*, data yang bisa diolah dalam perhitungan adalah data numerik yang berbentuk angka. Sedangkan data selain angka juga bisa diterapkan tetapi terlebih dahulu harus dilakukan pengkodean untuk mempermudah perhitungan jarak/kesamaan karakteristik yg dimiliki dari setiap objek. Setiap objek dihitung kedekatan jaraknya berdasarkan karakter yang dimiliki dengan pusat *cluster* yang sudah ditentukan sebelumnya, jarak terkecil antara objek dengan masing-masing *cluster* merupakan anggota *cluster* yang terdekat. Setelah jumlah *cluster* ditentukan, selanjutnya dipilih sebanyak 3 objek secara acak sesuai jumlah *cluster* yang dibentuk sebagai pusat *cluster* awal untuk dihitung jarak kedekatannya terhadap semua objek yang ada. Berhubung proses *iterasi* ini tidak dapat dipastikan jumlah perulangannya, untuk objek yang berjumlah besar perhitungan ini bisa dipermudah dengan menggunakan *software SPSS* (S*tatistical Package for Social Science*) Versi 17.0 yaitu dengan bantuan menu **Analyze** dan submenu **Classify** lalu pilih ***K-Means Cluster.***

PENUTUP

Berdasarkan uraian dari penulisan ini dapat diambil kesimpulan sebagai berikut:

1. Metode *K-Means Cluster Analysis* cukup efektif diterapkan dalam proses pengklasifikasian karakteristik terhadap objek penelitian. Algoritma *K-Means* juga tidak terpengaruh terhadap urutan objek yang digunakan, hal ini dibuktikan ketika penulis mencoba menentukan secara acak titik awal pusat *cluster* dari salah satu objek pada permulaan perhitungan. Jumlah keanggotaan *cluster* yang dihasilkan berjumlah sama ketika menggunakan objek yang lain sebagai titik awal pusat *cluster* tersebut. Namun, hal ini hanya berpengaruh pada jumlah *iterasi* yang dilakukan.
2. Algoritma *K-Means Cluster Analysis* pada dasarnya dapat diterapkan pada permasalahan dalam memahami perilaku konsumen, mengidentifikasi peluang produk baru dipasaran dan algoritma *K-Means* ini juga dapat digunakan untuk meringkas objek dari jumlah besar sehingga lebih memudahkan untuk mendiskripsikan sifat-sifat atau karakteristik dari masing-masing kelompok.

DAFTAR PUSTAKA

[1]. Simamora B. *Analisis Multivariat Pemasaran*. Jakarta: PT. Gramedia Pustaka Utama; 2005.
[2]. Agusta Y. K-Means-Penerapan, Permasalahan dan Metode Terkait. Denpasar, Bali: *Jurnal Sistem dan Informatika (Februari 2007) Vol. 3: 47-60;* 2007.
[3]. Rismawan dan Kusumadewi. Aplikasi K-Means untuk *Pengelompokan* Mahasiswa Berdasarkan Nilai Body Mass index (BMI) & Ukuran Kerangka. Yogyakarta*: Seminar Nasional Aplikasi Teknologi Informasi 2008 (SNATI 2008) ISSN: 1907-5022;* 2008.
[4]. Singarimbun M dan Effendi S. *Metode Penelitian Survai.* Ed rev. Jakarta: LP3ES; 1989.
[5]. Kusnandar D. *Metode Statistik dan aplikasinya dengan Minitab dan Excel*. Yogyakarta: Madyan Press, 2003.
[6]. Agusta Y. K-Means-Penerapan, Permasalahan dan Metode Terkait. Denpasar, Bali: *Jurnal Sistem dan Informatika (Februari 2007) Vol. 3: 47-60;* 2007.

EDIYANTO : FMIPA UNTAN, Jl. A Yani, ide.smart_yanto@yahoo.com
MUHLASAH NOVITASARI MARA : FMIPA UNTAN, Jl. A Yani, noveemara@gmail.com
NEVA SATYAHADEWI : FMIPA UNTAN, Jl. A Yani , neva_s04@yahoo.co.id