Analisis *Hybrid Mutual Clustering* menggunakan Jarak *Square Euclidean*

Astrid Alfira[1], Fariani Hermin[2], Eti Dwi Wiraningsih[3]
Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta
Jl. Rawamangun Muka Jakarta Timur 13320
E-mail: vinnaangelaw@gmail.com

ABSTRAK

Analisis kelompok berguna untuk mengelompokkan objek berdasarkan ukuran kemiripan, dimana konsep dasar dari analisis kelompok adalah pengukuran jarak dan kesamaan. Pengelompokan objek di dalam analisis kelompok dapat dilakukan dengan metode *bottom-up*, *top-down*, dan *Hybrid Mutual Clustering*. Pengelompokan objek dengan *bottom-up* menggunakan metode pengelompokan yang dimulai dari kelompok kecil menjadi kelompok yang lebih besar, pengelompokan objek dengan *top-down* menggunakan metode sebaliknya yaitu pengelompokan dengan memecah kelompok besar menjadi kelompok yang lebih kecil. Metode *Hybrid Mutual Clustering* baru diperkenalkan pada tahun 2006 oleh Hugh Chipman dan Robert Tibshirani, dimana metode ini mengkombinasikan kelebihan metode *bottom-up* dan *top-down*. Metode *Hybrid Mutual Clustering* yang digunakan dalam skripsi ini adalah metode pengelompokan *hybrid* menggunakan jarak *Square Euclidean* sebagai metode perhitungan jarak objek satu ke objek lainnya. Pemilihan kelompok terbaik dipilih berdasarkan nilai proporsi terbesar pada variabel dan perbedaan karakteristik antar variabel. Pada skripsi ini proporsi terbesar didapat darivariabel umur di bawah 14 tahun di kota Nusa Tenggara Barat dan Papua.

Kata kunci : Klaster, *Hybrid Mutual Clustering*, *Bottom-Up*, *Top-Down*, Jarak *Square Euclidean*.

I PENDAHULUAN

A. Latar Belakang

Cluster atau 'klaster' dapat diartikan kelompok; dengan demikian, pada dasarnya analisis klaster akan menghasilkan sejumlah klaster (kelompok). Analisis ini diawali dengan pemahaman bahwa sejumlah data tertentu sebenarnya mempunyai kemiripan diantara anggotanya. Karena itu, dimungkinkan untuk mengelompokkan anggota-anggota yang 'mirip' atau mempunyai karakteristik yang serupa tersebut dalam satu atau lebih dari satu klaster.

Analisis kelompok berguna untuk mengelompokkan objek berdasarkan ukuran kemiripan, dimana konsep dasar dari analisis kelompok adalah pengukuran jarak dan kesamaan. Pengelompokan objek di dalam analisis kelompok dapat dilakukan dengan metode *bottom-up*, *top-down*, dan *Hybrid Mutual Clustering*. Pengelompokan objek dengan *bottom-up* menggunakan metode pengelompokan yang dimulai dari kelompok kecil menjadi kelompok yang lebih besar, pengelompokan objek dengan *top-down* menggunakan metode sebaliknya yaitu pengelompokan dengan memecah kelompok besar menjadi kelompok yang lebih kecil. Metode *Hybrid Mutual Clustering* baru diperkenalkan pada tahun 2006 oleh Hugh Chipman dan Robert

Tibshirani, dimana metode ini mengkombinasikan kelebihan metode *bottom-up* dan *top-down*. Algoritma *bottom-up* baik dalam mengelompokkan ukuran sampel kecil dan sebaliknya, algoritma *top-down* baik dalam mengelompokkan ukuran sampel besar. Metode *Hybrid Mutual Clustering* yang digunakan adalah metode pengelompokan *hybrid* melalui *mutual cluster*. *Mutual cluster* adalah pengelompokan yang menggunakan jarak terbesar antara pasangan dalam kelompok yang lebih kecil dari jarak terpendek ke setiap titik di luar kelompok.

Dalam penelitian ini, metode yang dipakai menggunakan jarak *Square Euclidean* yang merupakan pengembangan dari jarak *Euclidean*. Sebagaimana namanya, *Square Euclidean* adalah ukuran jarak dengan mengkuadratkan selisih antara dua objek yang sama pada kelompok yang berbeda. Relatif untuk beberapa persoalan terutama menyangkut persoalan lokasi objek diselesaikan dengan penerapan *Square Euclidean*.

II LANDASAN TEORI

- **Analisis Kelompok**

Analisis kelompok merupakan suatu analisis multivariat yang digunakan untuk mengelompokan objek pengamatan menjadi beberapa kelompok berdasarkan ukuran kemiripan antar objek, sehingga objek-objek yang berada dalam satu kelompok memiliki kemiripan yang lebih homogen dibandingkan objek dari kelompok yang berbeda (Johnson & Wichern, 2002). Seperti diketahui, analisis kelompok akan membagi sejumlah data pada satu atau beberapa *cluster* tertentu. Sebuah *cluster* yang baik adalah cluster yang mempunyai ciri sebagai berikut:

- Homogenitas (kesamaan) yang tinggi antara anggota dalam satu *cluster* atau biasa disebut *within cluster*.
- Heterogenitas (perbedaan) yang tinggi antara cluster satu dengan *cluster* yang lain atau biasa disebut *between cluster*.

Proses pengolahan data sehingga sekumpulan data mentah dapat dikelompokkan menjadi satu atau beberapa *cluster* adalah sebagai berikut:

- Menetapkan ukuran jarak antar data. Mengukur kesamaan antar objek (*similarity*).
- Melakukan proses standardisasi data jika diperlukan.
- Melakukan proses *clustering*. Setelah data yang dianggap mempunyai satuan yang sangat berbeda sudah diseragamkan, langkah selanjutnya adalah membuat *cluster*. Proses inti dari *clustering* adalah pengelompokan data yang bisa dilakukan dengan 2 metode, yaitu metode hirarki dan metode non hirarki.

Lalu terdapat pula asumsi pada analisis kelompok diantaranya adalah sampel yang diambil benar-benar bisa mewakili populasi yang ada dan kemungkinan adanya korelasi antar objek. Jika terdapat korelasi, maka dianjurkan untuk melakukan analisis komponen utama yang akan dijelaskan pada subbab berikutnya.

- **Analisis Korelasi**

Analisis korelasi mencoba mengukur keeratan hubungan antara dua peubah melalui sebuah bilangan yang disebut koefisien korelasi. Ukuran hubungan linear antara dua peubah diduga dengan koefisien korelasi dirumuskan sebagai berikut (Walpole, 1995):

$$r = \frac{n\sum_{i=1}^{n} X_{1i}X_{2i} - \left(\sum_{i=1}^{n} X_{1i}\right)\left(\sum_{i=1}^{n} X_{2i}\right)}{\sqrt{\left[n\sum_{i=1}^{n} X_{1i}^2 - \left(\sum_{i=1}^{n} X_{1i}\right)^2\right]\left[n\sum_{i=1}^{n} X_{2i}^2 - \left(\sum_{i=1}^{n} X_{2i}\right)^2\right]}}.$$

- **Analisis Komponen Utama**

Analisis Komponen Utama (Principal Component Analysis) adalah analisis multivariat yang mentransformasi variabel-variabel asal yang saling berkorelasi menjadi variabel-variabel baru yang tidak saling berkorelasi dengan mereduksi sejumlah variabel tersebut sehingga mempunyai dimensi yang lebih kecil namun dapat menerangkan sebagian besar keragaman variabel aslinya. Banyaknya komponen utama yang terbentuk sama dengan banyaknya variabel asli. Pereduksian (penyederhanaan) dimensi dilakukan dengan kriteria persentase keragaman data yang diterangkan oleh beberapa komponen utama pertama. Apabila beberapa komponen utama pertama telah menerangkan lebih dari 75 % keragaman data asli, maka analisis cukup dilakukan sampai dengan komponen utama tersebut.

- **Mutual Cluster**

Mutual cluster adalah suatu pengelompokan yang menggunakan jarak terbesar antara pasangan dalam kelompok yang lebih kecil dari jarak terpendek ke setiap titik di luar kelompok. Hal ini berarti bahwa jarak maksimal antar obyek dalam sebuah mutual cluster lebih kecil dibandingkan jarak minimal beberapa obyek di luar mutual cluster. Data yang terkandung dalam sebuah mutual cluster tidak pernah dipisahkan (Chipman dan Tibshirani, 2006). Metode tersebut memiliki beberapa implikasi dalam sebuah mutual cluster. Implikasi yang paling jelas adalah untuk mendukung gagasan bahwa dalam sebuah mutual cluster berisi informasi pengelompokan yang kuat, tidak peduli pendekatan linkage mana yang digunakan. Hal ini dapat membantu dalam interpretasi metode bottom-up. Informasi tambahan tesebut dapat membantu dalam interpretasi dari mutual cluster, atau dalam menentukan keputusan untuk pembagian kelompok. Hybrid ini juga mempertahankan metode top-down yang akurat membagi data menjadi pengelompokan yang baik.

- **Metode Pengelompokan Objek**
  1. Metode Pengelompokan Hirarki

- Single Linkage
  Pengambilan jarak berdasarkan jarak minimum.
- Complete Linkage
  Pengambilan jarak berdasarkan jarak maksimum.
- Average Linkage
  Pengambilan jarak berdasarkan jarak rata-rata.

2. Metode Pengelompokan Non-Hirarki
   Metode Pengelompokkan hirarki digunakan apabila belum ada informasi jumlah kelompok. Sedangkan metode pengelompokkan nonhirarki bertujuan pengelompokkan n objek ke dalam k kelompok (k < n). Salah satu pengelompokkan pada non hirarki adalah dengan menggunakan metode K-Means.

- **Jarak Square Euclidean**

  Jarak Square Euclidean merupakan jarak yang dikembangkan dari jarak Euclidean. Pada jarak Euclidean, jarak tersebut mempunyai tiga asumsi yang diantaranya adalah antar peubah tidak saling berkorelasi, memiliki satuan pengukuran yang sama, dan pengukuran pembakuan mempunyai rata-rata nol dan standar deviasi satu. ormula jarak Square Euclidean adalah sebagai berikut (Hair dkk., 2010):

$$d_{(X_i,X_j)} = \sum_{q=1}^{p} (X_{iq} - X_{jq})^2.$$

- **Pengelompokan Metode Bottom-Up**

Pengelompokan dengan menggunakan metode bottom-up adalah suatu metode hierarki dimana n buah kelompok digabungkan menjadi satu kelompok tunggal. Metode bottom-up ini meletakkan setiap objek data sebagai sebuah kelompok tersendiri (atomic cluster) yang selanjutnya kelompok-kelompok tersebut bergabung menjadi kelompok besar sampai akhirnya semua objek menyatu dalam sebuah kelompok tunggal. Jarak antar objek diperlukan pada tahap awal dalam penggabungan 2 kelompok dengan metode agglomerative (Hair dkk., 2010).

Algoritma bottom-up dimulai dari menginput data amatan Xij yang kemudian dihitung pusat objek amatan (dij). Selanjutnya buat matriks dij yang kemudian dilihat apakah nilai dij minumum, ulangi hitung nilai pusat objek amatan apabila nilai dij belum minimum. Setelah dapat nilai dij minimum, gabungkan objek amatan yang sama menjadi satu kelompok. Output berupa k kelompok objek amatan.

- **Pengelompokan Metode Top-Down**

Pengelompokan dengan metode top-down adalah membagi n objek ke dalam k kelompok yang bertujuan untuk mengelompokkan objek sehingga jarak antar objek ke pusat kelompok di dalam satu kelompok minimum.

Proses pertama dalam mengelompokkan dengan menggunakan metode top-down yang bersifat non-hierarki (k-means) adalah terdapat data amatan Xi dan Xj . Kemudian, partisikan obyek ke dalam k kelompok. Langkah selanjutnya, hitung pusat kelompok dimana pusat kelompok itu sendiri merupakan rata-rata dari keseluruhan obyek yang berasa dalam kelompok tersebut. Setelah itu, hitung jarak setiap obyek ke pusat kelompok dengan menggunakan jarak Square Euclidean. Jika terdapat obyek yang berpindah dari posisi awal, maka pusat kelompok dihitung kembali dan periksa kembali posisi obyek. Ulangi langkah-langkah tersebut sampai tidak ada obyek yang berpindah posisi. Perhitungan berhenti ketika obyek sudah tidak berpindah posisi dan membentuk kelompok Hybrid Mutual Clustering.

III PEMBAHASAN

A. Pengelompokan Menggunakan Metode Hybrid Mutual Clustering dengan Jarak Square Euclidean

Metode Hybrid Mutual Clustering merupakan metode penggabungan antara metode bottom-up dan top-down. Pengelompokan Hybrid Mutual Clustering ini menggunakan jarak Square Euclidean. Hybrid mutual disini mengkombinasikan kelebihan metode bottom-up clustering (agglomerative) dan top-down clustering (k-means). Metode Hybrid terdiri dari dua metode yaitu bottom-up yaitu metode pengelompokan dimulai dari kelompok kecil menjadi kelompok yang lebih besar (agglomerative) dan top-down yaitu metode pengelompokan dengan memecah kelompok besar menjadi kelompok lebih kecil seperti metode k-means membagi sebanyak k kelompok.

Berikut akan dijelaskan algoritma pengelompokkan Hybrid Mutual Clustering menggunakan jarak Square Euclidean:

- Masukkan data amatan Xij dengan i = 1, ..., n dan j = 1, ..., n.
- Hitung dij untuk mencari jarak pusat antar objek amatan, dengan persamaan berikut:

$$d_{(X_i,X_j)} = \sum_{q=1}^{p} (X_{iq} - X_{jq})^2.$$

- Bentuk matriks dij .
- Tentukan nilai minimum dij . Jika dij minimum maka proses berlanjut ke tahap selanjutnya. Namun, jika dij tidak minimum, maka proses mengulang dari tahap menghitung jarak pusat.
- Gabungkan 2 data amatan yang jaraknya paling minimum ke dalam satu cluster, dimana dij=0.
- Bentuk sebanyak k kelompok objek amatan.

$$c_{kj} = \frac{\sum_{j-1}^{p} x_{ij}}{p}$$

- Tentukan pusat dengan peubah sebanyak p.
- Hitung:

$$d_{ik} = \sum_{j-1}^{p} (x_{ij} - c_{kj})^2.$$

- Kelompokkan berdasarkan minimum dik.

- Tentukan nilai ckj. Jika ckj berubah maka proses berlanjut ke tahap selanjutnya. Jika ckj tidak berubah, maka proses mengulang dari menghitung dik.
- Pusat ckj berubah.
- Identifikasi data amatan. Jika data amatan berubah posisi maka proses mengulang dari tahap 7. Jika data amatan tidak berubah, maka proses berlanjut ke tahap 13.
- Didapat Cluster tetap yang memiliki proporsi terbaik.

B. Penerapan Metode Hybrid Mutual Clustering dengan Jarak Square Euclidean ke dalam contoh kasus.

Sumber data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari Badan Pusat Statistik (BPS) dengan judul Persentase Penduduk Buta Aksara Menurut Provinsi di Indonesia. Pada situs tersebut data yang tertera ada dalam periode 2003-2015, tetapi yang diambil oleh penulis hanya tahun 2015 saja. Unit pengamatan yang dipakai ada 34 provinsi di Indonesia, meliputi Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, dan Papua. Data yang dipakai hanya mencakup 3 variabel, yaitu penduduk dengan umur di bawah 14 tahun, penduduk dengan umur 15 sampai 44 tahun, dan penduduk dengan umur di atas 45 tahun. Variabel tersebut digunakan karna tersediaan data.

Metode Hybrid Mutual Clustering ini akan diaplikasikan ke dalam pengelompokan Penduduk Buta Aksara Menurut Provinsi di Indonesia tahun 2015 menggunakan software SPSS. Setelah melakukan Hybrid Mutual Clustering, dihasilkan output dengan penjelasan sebagai berikut:

- Tabel output Proximity Matrix yang terdapat pada lampiran. Pada tabel tersebut menunjukkan bahwa semua data sejumlah 34 obyek telah diproses tanpa ada data yang hilang. Tabel tersebut menunjukkan matriks jarak antara variabel satu dengan variabel yang lain. Semakin kecil jarak Euclidean, maka semakin mirip kedua variabel tersebut sehingga akan membentuk kelompok (cluster).
- Selain tabel Proximity Matrix, terdapat pula tabel outpyt yang berjudul Agglomeration Schedule. Tabel tersebut merupakan hasil proses clustering dengan metode Between Group Linkage. Setelah jarak antar variabel diukur dengan jarak Square Euclidean, maka dilakukan pengelompokan, yang dilakukan secara bertingkat.
- Terdapat satu lagi hasil output SPSS yaitu dendogram. Dendogram berguna untuk menunjukkan anggota cluster yang ada jika akan ditentukan berapa cluster yang seharusnya dibentuk.

IV PENUTUP

A. Kesimpulan

Penerapan metode Hybrid Mutual Clustering ini pada kasus Penduduk Buta Aksara di Indonesia tahun 2015 didapatkan hasil bahwa proporsi untuk komponen utama terbesar terdapat pada variabel umur di bawah 14 tahun. Proporsi tujuannya melihat variabel mana yang memiliki pengaruh lebih besar dari yang lainnya, porporsi juga memengaruhi jarak yang diperoleh dari setiap provinsi yang memiliki karakteristik yang sama terhadap variabel umur. Anggota sebuah cluster tentu mempunnyai kemiripan satu dengan yang lain, dan mereka tentu juga berbeda dengan anggota cluster yang lain. Pada kasus ini, terlihat kota Nusa Tenggara Barat dan Papua mempunyai karakteristik yang berbeda dengan kota lain.

B. Saran

1. Penerapan dari metode Hybrid Hierarchical Clustering dapat menggunakan jarak selain Square Euclidean, antara lain: jarak Pearson, jarak Mahalanobis
2. Dalam metode pengelompokan, data yang digunakan harus sesuai dengan karakteristik dalam metode tersebut, untuk mendapatkan hasil yang sesuai.

DAFTAR PUSTAKA

Agustina, Mitakda, dan Solimun. 2013. "Pemilihan Metode Pengelompokan Terbaik Kabupaten/Kota Berdasarkan Indikator Pendidikan Menggunakan Hybrid Melalui Mutual Cluster, Bottom-up, dan Top-down", Jurnal Mahasiswa Statistik Universitas Brawijaya-Malang Vol. 1 No. 3. Hal. 205-208.

Bikriyah. 2014. "Analisis Hybrid Hierarchical Clustering Melalui Mutual Cluster, Bottom-up, dan Top-down Menggunakan Jarak Euclidean dan Mahalanobis", Jurnal Mahasiswa Statistik Universitas Brawijaya-Malang Vol. 2 No. 5. Hal. 397-400.

Chipman dan Tibshirani. 2006. Hybrid Hierarchical Clustering With Applications To Microarray Data, Biostatistics Journal-Oxford England. Hal. 286-301.

Hair, dkk. 2010. Multivariate Data Analysis, Seventh Ed. New Jersey: Prentice Hall International, Inc.

Johnson dan Wichern. 2002. Applied Multivariate Analysis, Fifth Edition. New Jersey: Prentice Hall, Inc.

Manly. 1988. Multivariate Statistical Methods. New York: Chapman Hall.

Mariyani, dkk. 2011. "Penerapan Hybrid Hierarchical Clustering melalui Mutual Cluster dalam Pengelompokan Kabupaten di Jawa Timur berdasarkan Variabel Sektor Pertanian". Jurnal Mahasiswa Statistik FMIPA Institut Sepuluh November Surabaya.

Santoso. 2010. Statistik Multivariat. Jakarta: PT Elex Media Komputindo.

Walpole. 1995. Pengantar Statistika. Edisi ke-5. Jakarta: Terjemahan Bambang Sumantri, Gramedia.

-------. Data Penduduk Buta Aksara menurut Provinsi tahun 2015. [ON LINE] http://data.go.id/dataset/persentase-penduduk-buta-aksara/resource/3e89671a-3199-4716-ba05-b39e00162b54 (di-unduh pada tanggal 8 Oktober 2016, pukul 19:04)