RANCANG BANGUN ALGORITMA CLUSTERING ENSEMBEL FUZZY-K-PROTOTYPES-DPC UNTUK DATA BERTIPE CAMPURAN

Laila Qadrini

RANCANG BANGUN ALGORITMA CLUSTERING ENSEMBEL FUZZY-K-PROTOTYPES-DPC UNTUK DATA BERTIPE CAMPURAN

Laila Qadrini

RANCANG BANGUN ALGORITMA CLUSTERING ENSEMBEL FUZZY-K-PROTOTYPES-DPC UNTUK DATA BERTIPE CAMPURAN

Penulis:
Laila Qadrini

Editor:
Syifa Ismayanti

Layouter :
Tim Kreatif PRCI

Cover:
Rusli

Cetakan Pertama : Juni 2025

Perkumpulan Rumah Cemerlang Indonesia
ANGGOTA IKAPI JAWA BARAT
Pondok Karisma Residence Jalan Raflesia VI D.151
Panglayungan, Cipedes Tasikmalaya – 085223186009

Website : www.rcipress.rcipublisher.org
E-mail : rumahcemerlangindonesia@gmail.com

- Cet. I – : Perkumpulan Rumah Cemerlang Indonesia, 2025
; 14,8 x 21 cm
ISBN

KATA PENGANTAR

Puji syukur kami panjatkan ke hadirat Tuhan Yang Maha Esa yang telah memberikan rahmat serta karunia-Nya kepada kami, sehingga kami berhasil menyelesaikan buku dengan judul "Rancang Bangun Algoritma Clustering Ensembel Fuzzy-K-Prototypes-DPC untuk Data Bertipe Campuran" sesuai dengan yang ditargetkan.

Buku ini berisikan penjelasan mengenai teori dan implementasi algoritma pengelompokan data campuran dengan pendekatan *ensembel clustering*, yang menggabungkan metode Fuzzy, K-Prototypes, dan Density Peaks Clustering for Mixed data (DPC-M). Pembahasan dalam buku ini mencakup mulai dari konsep dasar data mining, penerapan dalam konteks beasiswa Bidikmisi, teknik preprocessing data, hingga analisis hasil pengelompokan dan evaluasi performa metode melalui indeks validasi internal maupun eksternal.

Buku ini diharapkan dapat menjadi referensi bagi mahasiswa, peneliti, maupun praktisi yang tertarik dalam bidang pengolahan data, khususnya pada permasalahan clustering data bertipe campuran. Kami menyadari bahwa buku ini masih jauh dari sempurna, oleh karena itu kritik dan saran dari semua pihak yang bersifat membangun selalu kami harapkan demi kesempurnaan buku ini di masa yang akan datang.

Akhir kata, kami sampaikan terima kasih kepada semua pihak yang telah berperan serta dalam penyusunan buku ini dari awal hingga akhir. Semoga Tuhan Yang Maha Esa senantiasa meridhoi segala usaha kita. Amin.

Juni 2025, Penulis

DAFTAR ISI

BAB I
PENGANTAR ALGORITMA PENGELOMPOKAN ENSEMBEL FUZZY, K-PROTOTYPES DAN DENSITY PEAKS CLUSTERING MIXED (DPC-M)

Analisis pengelompokan (Kelompok *Analysis*) merupakan salah satu metode dalam data Mining. Metode kelompok dalam data mining berbeda dengan metode konvensional yang biasa digunakan untuk pengelompokkan. Perbedaannya adalah data mining memiliki dimensi data yang tinggi yaitu bisa terdiri dari puluhan ribu atau jutaan *record* dengan puluhan ataupun ratusan atribut. Selain itu pada data mining data bisa terdiri dari tipe data campuran seperti data numerik dan kategorikal. Permasalahan yang sering ditemui dalam analisis pengelompokan adalah data yang berskala campuran numerik dan kategorik. Metode yang seringkali dilakukan untuk pengelompokan data berskala campuran adalah dengan mentransformasi data kategorik menjadi data numerik dan sebaliknya. Dewangan, Sharma, dan Akasapu (2010) melakukan transformasi variabel kategorik ke dalam bentuk numerik, kemudian pengelompokan objek dilakukan dengan metode pengelompokan data numerik. Kelebihan metode transformasi adalah dapat mengurangi kompleksitas dalam komputasi. Akan tetapi, metode tersebut memiliki kelemahan dalam menentukan transformasi yang tepat agar tidak kehilangan banyak informasi dari data aslinya.

Selain pengelompokan dengan metode transformasi tersebut, dikembangkan sebuah metode pengelompokan ensembel untuk data campuran oleh He, Xu dan Deng (2005). Pengelompokan ensembel (pengelompokan *ensemble*) adalah teknik pengelompokan untuk menggabungkan hasil pengelompokan beberapa algoritma pengelompokan untuk mendapatkan kelompok yang lebih baik (He, Xu, dan Deng, 2005a). He Xu, dan Deng (2005b) menerapkan teknik pengelompokan ensembel untuk mengelompokkan penyakit hati dan persetujuan pengajuan kartu kredit dengan Kelompok *Ensemble Based Mixed Data Clustering* (CEBMDC) dan mengombinasikan tahapan ensembel ke dalam algoritma pengelompokan untuk data kategorik, yaitu algoritma *Squeezer*. Algoritma *Squeezer* merupakan metode pengelompokan yang efektif digunakan untuk data yang berjumlah besar (Reddy & Kavitha, 2010). Alvionita (2017) melakukan perbandingan hasil antara metode ensembel ROCK dan ensembel SWFM. Kedua metode digunakan pada studi kasus pengelompokkan aksesi jeruk hasil fusi protoplasma yang merupakan data campuran numerik dan kategorik. Metode pengelompokan terbaik ditentukan dengan kriteria rasio antara simpangan baku di dalam kelompok (*SW*) dan simpangan baku antar kelompok (*SB*) terkecil.

Berdasarkan 25 objek pengamatan pada studi kasus, metode ensemble ROCK dengan nilai θ sebesar 0,27 menghasilkan tiga kelompok dengan nilai rasio sebesar 0,1358, sedangkan metode ensembel SWFM menghasilkan dua kelompok dengan nilai rasio sebesar 0,3059. Hasil tersebut menunjukkan bahwa metode ensembel ROCK memberikan hasil pengelompokan lebih baik daripada

metode ensembel SWFM, J. Suguna & M. Arul Selvi (2015) membagi dataset campuran asli menjadi kumpulan data numerik dan kumpulan data kategorik dan dikelompokkan menggunakan algoritma pengelompokan tradisional (K-*Means* and K-*Modes*) dan algoritma pengelompokan fuzzy (Fuzzy C-*Means* dan Fuzzy C-*Modes*). Kelompok yang dihasilkan dikombinasikan dengan metode ensemble pengelompokan dan dievaluasi dengan ukuran f-*measure* dan *entropy*. Ditemukan bahwa pembagian data lebih menguntungkan dan menerapkan algoritma pengelompokan *fuzzy* menghasilkan hasil yang lebih baik daripada algoritma pengelompokan tradisional. Algoritma - Prototypes diajukan oleh Huang pada tahun 1997, dan didasarkan pada gagasan algoritma -*means*. Metode ini menggabungkan pusat kelompok dari atribut numerik dan atribut modus kategorik untuk dibangun sebuah pusat campuran baru. Pusat data adalah prototipe, dan ukuran jarak jauh proses pengelompokan -means digunakan untuk mengelompokkan data campuran secara langsung. Rani Nooraeni (2015) telah melakukan simulasi yang hasilnya, metode K-*prototype* dapat meningkatkan kehomogenan dalam kelompok dan ketidakmiripan maksimal antar kelompok, data yang digunakan adalah Dataset Podes 2011.

Ukuran data yang digunakan pada simulasi ini sebanyak 77.961 objek atau desa dan 37 atribut yang terdiri dari 17 atribut numerik dan 20 atribut kategorikal. Pada data numerik dinormalisasi terlebih dahulu. Jumlah kluster yang dibentuk adalah 10 kelompok. Reddy & Kavitha (2012) juga membandingkan pengelompokan antara metode ensembel dan metode *K-Prototype*. Hasil yang diperoleh adalah metode ensembel memberikan rata-rata kesalahan lebih rendah

daripada metode *KPrototype.* Liu et al (2017) melakukan penelitian dengan mengacu pada data campuran yang terdiri dari atribut numerik dan kategorik, mengajukan ukuran disimilaritas baru dan algoritma pengelompokan baru. Hasil penelitian menunjukkan bahwa metode baru pengelompokan ini digabung data dengan pencarian cepat dan menemukan puncak kepadatan (DPC-M) terbukti layak dan efektif pada dataset UCI. Namun, penelitian-penelitian tersebut tidak dilakukan perbandingan hasil pengelompokan antara ketiga metode. Oleh karena itu, pada penelitian ini dilakukan perbandingan hasil pengelompokan antara metode pengelompokan ensembel *Fuzzy*, K-*prototypes* dan DPC-M. Hasil pengelompokan ketiga metode tersebut dilihat berdasarkan dua tipe untuk mengukur validitas kelompok, yaitu ukuran eksternal dan ukuran internal (Steinbach, Karypis, dan Kumar, 2000).

BAB II
DATA MINING PADA PENGELOLAAN BEASISWA BIDIK MISI

A. Tujuan Bidikmisi

1. Meningkatkan akses dan kesempatan belajar di perguruan tinggi bagi peserta didik yang tidak mampu secara ekonomi namun memiliki prestasi akademik yang baik
2. Meningkatkan prestasi mahasiswa, baik pada bidang kurikuler, kokurikuler, maupun ekstrakurikuler
3. Menimbulkan dampak iring bagi mahasiswa dan calon mahasiswa lain untuk berkarakter dan selalu meningkatkan prestasi; melahirkan lulusan yang mandiri, produktif, dan memiliki kepedulian sosial sehingga mampu berperan dalam upaya pemutusan mata rantai kemiskinan dan pemberdayaan masyarakat.

B. Syarat penerima Bidik misi

Penerima Bidikmisi memiliki syarat sebagai berikut.

1. Pendapatan kotor orang tua/wali gabungan (suami dan istri) setinggi-tingginya Rp4.000.000,00 (empat juta rupiah) atau pendapatan kotor gabungan orang tua/wali dibagi jumlah anggota keluarga maksimal Rp750.000,00 (tujuh ratus lima puluh ribu rupiah)
2. Ditetapkan oleh perguruan tinggi setiap tahun akademik
3. Mahasiswa aktif dan sedang menjalani perkuliahan pada semester normal

C. Jangka Waktu Pemberian Bantuan

Jangka waktu pemberian Bidikmisi diberikan dengan ketentuan sebagai berikut.

1. Sampai dengan semester 8 (delapan) untuk S1/D4
2. Sampai dengan semester 6 (enam) untuk D3
3. Sampai dengan semester 4 (empat) untuk D2
4. Sampai dengan semester 2 (dua) untuk D1

Penerima Bidikmisi yang cuti diberhentikan bantuannya. Pengelola perguruan tinggi dapat merekomendasikan yang bersangkutan menerima Bidikmisi pada saat aktif kembali. Keputusan akhir pengaktifan diputuskan oleh Pengelola Pusat.

D. Hak dan Kewajiban Penerima Bidikmisi

Hak dan kewajiban penerima Bidikmisi adalah sebagai berikut.

1. Hak

a. Mendapatkan akses dan kesempatan mendapatkan pendidikan yang berkualitas sama dengan peserta didik lain di Perguruan Tinggi Penyelenggara Bidikmisi.
b. Wajib mendapatkan pembebasan biaya yang terdiri atas:
    1) UKT/SPP atau sejenisnya yang bersifat operasional pendidikan
    2) Biaya awal pendidikan yang mencakup biaya gedung, pembinaan, investasi, infak atau sejenisnya
    3) Biaya praktikum di laboratorium, bahan, atau biaya pendidikan lain yang belum dicakup UKT/SPP

4) Biaya yudisium

c. Mendapatkan pembebasan biaya pendidikan sesuai jangka waktu pemberian bantuan
d. Mendapatkan biaya hidup sekecil kecilnya Rp650.000,00 (enam ratus lima puluh ribu rupiah) per bulan yang akan dibayarkan 6 (enam) bulan sekali
e. Mendapatkan pembinaan dan fasilitasi dari perguruan tinggi pengelola untuk menunjang kegiatan akademik dan kemahasiswaan untuk mewujudkan misi program.

2. Kewajiban

a. Menjunjung tinggi negara kesatuan Republik Indonesia dengan dasar negara Pancasila dan UUD 1945.
b. Memenuhi kontrak hasil Bidikmisi dengan Perguruan Tinggi Penyelenggara, termasuk namun tidak terbatas pada kewajiban akademis dan administratif.
c. Berperan aktif dan berkontribusi dalam pelaksanaan Tridarma Perguruan Tinggi.

E. Mekanisme Pendaftaran Bidikmisi

1. Pendaftaran Daring (*On line*)

Tata cara pendaftaran Bidikmisi melalui SNMPTN, SBMPTN, PMDK Politeknik atau Seleksi Mandiri Perguruan Tinggi secara *online* pada laman Bidikmisi (http://bidikmisi.belmawa.ristekdikti.go.id/) adalah sebagai berikut.

a. Tahapan pendaftaran Bidikmisi
    1) Sekolah mendaftarkan diri sebagai institusi pemberi rekomendasi ke laman Bidikmisi dengan melampirkan hasil pindaian (*scan*) (Lampiran 1 bagian persetujuan dan tanda tangan) untuk mendapatkan nomor Kode Akses Sekolah
    2) Direktorat Jenderal Pembelajaran dan Kemahasiswaan memverifikasi pendaftaran dalam kurun waktu 1 x 24 jam pada hari dan jam kerja
    3) Sekolah merekomendasikan masing-masing siswa melalui laman Bidikmisi menggunakan kombinasi NPSN dan kode akses yang telah diverifikasi. Sekolah memberikan nomor pendaftaran dan kode akses kepada masing-masing siswa yang sudah direkomendasikan
    4) Siswa mendaftar melalui laman Bidikmisi dan menyelesaikan semua tahapan yang diminta di dalam sistem pendaftaran.
b. Siswa yang sudah menyelesaikan pendaftaran Bidikmisi mendaftar seleksi nasional atau mandiri yang telah diperoleh sesuai ketentuan masing-masing pola seleksi melalui alamat berikut:
    1) SNMPTN melalui http://www.snmptn.ac.id
    2) SBMPTN melalui http://www.sbmptn.ac.id
    3) PMDK Politeknik melalui http://pmdk.politeknik.or.id
    4) Seleksi Mandiri PTN sesuai ketentuan masing-masing PTN
    5) Seleksi Mandiri PTS sesuai ketentuan masing masing PTS.

c. Siswa yang mendaftar dan ditentukan lolos melalui seleksi masuk, melengkapi berkas, dan berkas dibawa pada saat pendaftaran ulang, yaitu:
    1) Kartu peserta dan formulir pendaftaran program Bidikmisi yang dicetak dari laman Bidikmisi
    2) Kartu Indonesia Pintar (KIP), atau bantuan pemerintah sejenis lainnya (jika ada)
    3) Siswa yang belum memenuhi syarat butir (b) di atas, harus membawa Surat Keterangan Penghasilan Orang Tua/Wali atau Surat Keterangan Tidak Mampu yang dapat dibuktikan kebenarannya, yang dikeluarkan oleh Kepala Desa/Kepala Dusun/Instansi tempat orang tua bekerja/tokoh masyarakat
    4) Fotokopi Kartu Keluarga atau Surat Keterangan tentang Susunan Keluarga
    5) Fotokopi rekening listrik bulan terakhir (apabila tersedia aliran listrik) dan/ atau bukti pembayaran PBB (apabila mempunyai bukti pembayaran) dari orang tua/wali-nya
    6) Berkas pendukung lainnya yang diminta oleh perguruan tinggi dan Kopertis.

F. Mekanisme Penetapan

Bagi calon mahasiswa penerima Bidikmisi yang telah dinyatakan diterima di Perguruan Tinggi, akan dilakukan hal-hal sebagai berikut:

1. Verifikasi kelayakan penerima Bidikmisi oleh perguruan tinggi dan Kopertis;
2. Penetapan mahasiswa penerima Bidikmisi oleh perguruan tinggi dan Kopertis.

G. Data Mining

Data mining berisi pencarian *trend* atau pola yang diinginkan dalam database yang besar untuk membantu pengambilan keputusan diwaktu yang akan datang. Harapannya, perangkat data mining mampu mengenali pola-pola ini dalam data dengan masukan yang minimal. Pola-pola ini dikenali oleh perangkat tertentu yang dapat memberikan suatu analisis data yang berguna dan berwawasan yang kemudian dapat dipelajari lebih teliti, yang mungkin saja menggunakan perangkat pendukung keputusan yang lainnya (Alith, 2016). Definisi sederhana dari data mining adalah ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di *database* yang besar. Data mining adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar. Hasil dari pengolahan data dengan menggunakan metode data mining ini dapat digunakan untuk mengambil keputusan di masa depan.

Berdasarkan penjelasan dari beberapa penelitian yang telah dilakukan sebelumnya, maka dapat disimpulkan bahwa hasil dari proses *data mining* haruslah informasi yang baru dan bermanfaat untuk keperluan kedepannya serta mudah dimengerti. Bidikmisi merupakan program pemerintah untuk memberikan akses pendidikan tinggi kepada masyarakat miskin untuk dapat memutus mata rantai kemiskinan. Sampai saat ini jumlah penerima Bidikmisi sudah mencapai angka 432.409 mahasiswa, sehingga berkontribusi untuk meningkatkan Angka Partisipasi Kasar (APK) Pendidikan Tinggi. Bidikmisi juga memiliki skema yang berbeda dengan bantuan biaya

pendidikan lain, dengan filosofinya untuk menjemput penerima.

Bidikmisi memberikan jaminan pembiayaan mulai dari pendaftaran sampai penerima Bidikmisi menuntaskan pendidikan tinggi. Bidikmisi adalah bantuan biaya pendidikan dari Kementerian Riset Teknologi dan Pendidikan Tinggi Republik Indonesia yang memberikan fasilitas pembebasan biaya pendidikan dan subsidi biaya hidup. Bidikmisi diberikan kepada penerima selama 8 (delapan) semester untuk S1 / D4, 6 (enam) semester untuk D3, 4 (empat) semester untuk D2, dan 2 (dua) semester untuk D1. Besaran subsidi biaya hidup yang diberikan serendah-rendahnya Rp650.000,00 per bulan diberikan setiap 6 bulan. Adapun pembebasan biaya pendidikan mencakup semua biaya yang dibayarkan ke Perguruan Tinggi untuk kepentingan pendidikan.

H. Data Numerik dan Data Kategorik

1. Numerik
   Nilai numerik termasuk nilai real (pecahan) dan integer (bilangan bulat). Fitur dengan nilai numerik memiliki 2 properti penting, yaitu: setiap nilai memiliki urutan dan memiliki relasi jarak.
2. Kategorik
   Dinyatakan dengan sama dengan atau tidak sama dengan, variabel kategori yang memiliki 2 nilai dapat dikonversi menjadi variabel numerik dengan 2 nilai values (0 atau 1). Variabel pengkodean dengan N buah nilai dapat dikonversikan ke dalam N buah variabel bertipe numerik yang memiliki nilai biner untuk setiap kategorikal. Pengkodean ini disebut "*dummy variabels*".

BAB III
ANALISIS DATA MINING

A. Analisis Kelompok

Analisis kelompok merupakan salah satu teknik data mining yang bertujuan untuk mengidentifikasi sekelompok obyek yang mempunyai kemiripan karakteristik tertentu yang dapat dipisahkan dengan kelompok obyek lainnya, sehingga obyek yang berada dalam kelompok yang sama relatif lebih homogen daripada obyek yang berada pada kelompok yang berbeda. Jumlah kelompok yang dapat diidentifikasi tergantung pada banyak dan variasi data obyek.

Analisis ini mempunyai tujuan utama untuk mengelompokkan objek-objek pengamatan menjadi beberapa kelompok berdasarkan karakteristik yang dimilikinya. Analisis kelompok mengelompokkan objek-objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam kelompok yang sama, serta mempunyai kemiripan satu dengan yang lain. (Johnson & Wichern, 2007). Beberapa manfaat analisis kelompok adalah eksplorasi data variabel ganda, reduksi data, dan prediksi keadaan objek. Hasil analisis kelompok dipengaruhi oleh objek yang dikelompokkan, variabel yang diamati, ukuran kemiripan atau ketakmiripan yang digunakan, skala ukuran yang digunakan, serta metode pengelompokan yang digunakan.

Ukuran kemiripan dan ketidakmiripan merupakan hal yang sangat mendasar dalam kelompok analisis. Algoritma pengelompokan menggunakan ukuran kemiripan atau ketidakmiripan digunakan untuk menggabungkan atau memisahkan data objek dari suatu data. Ukuran kemiripan biasanya digunakan oleh algoritma pengelompokan untuk menganalisis data kategori, sedangkan ukuran ketidakmiripan digunakan oleh algoritma pengelompokan untuk menganalisis data numerik. Ukuran ketakmiripan antara objek ke-*i* dengan objek ke-*j* (d_{ij}) merupakan fungsi yang memiliki sifat-sifat sebagai berikut $d_{ij} 0$, $d_{ii} = 0$, $d_{ij} = d_{ji}$, dan $d_{ik} + d_{jk}\ d_{ij}$, untuk setiap *i, j* dan *k*. Semakin besar nilai ukuran ketakmiripan antara dua objek maka semakin besar pula perbedaan antara kedua objek tersebut, sehingga makin cenderung untuk tidak berada dalam kelompok yang sama. (Johnson & Wichern, 2007).

Ukuran kemiripan dan ketakmiripan pada umumnya diukur berdasarkan jarak. Salah satu faktor yang sangat berpengaruh terhadap hasil dari kelompok yang dibentuk adalah jarak antar objek pengamatan (Sharma, 1996). Oleh karena itu, dibutuhkan suatu alat ukur untuk menentukan jarak antar objek pengamatan. Berikut ini merupakan metode-metode pengukuran jarak antara objek ke-*i* (X_i) dengan objek ke-*j* (X_j)berdasarkan karakteristivariabel yang dikelompokkan.

1. Metode pengukuran jarak untuk variabel kategorik biner bila variabel yang diamati berupa variabel biner yang hanya memiliki dua macam karakter yang berbeda (0,1), maka variabel yang diamati dapat dibentuk suatu tabel kontingensi seperti ditunjukkan pada Tabel 1 Perhitungan ukuran jarak antara variabel dan untuk

pengukuran data biner dapat menggunakan beberapa ukuran yang disajikan pada Tabel 2

Tabel 1 Tabel kontingensi data biner

Kategori X_i	Kategori X_j		Total
	1	0	
1	a	b	a+b
0	c	d	c+d

Tabel 2 Ukuran jarak data biner

Jenis	Rumus
Russel dan Rao	$RR(x_i, x_j) = \frac{a}{a+b+c+d}$
Simple matching	$SM(x_i, x_j) = \frac{a+d}{a+b+c+d}$
Jaccard	$JACCARD(x_i, x_j) = \frac{a}{a+b+c}$
Dice Czekanowski Sorensen	$DICE(x_i, x_j) = \frac{2a}{2a+b+c}$

2. Metode pengukuran jarak untuk variabel kategorik nominal Pada pengamatan dengan variabel nominal maka pengukuran memiliki konsep yang sama dengan *simple matcing coefficient* maupun *dice,* dimana kategorinya dapat lebih dari dua macam. Dengan jumlah variabel sebanyak *m,* maka rumus untuk pengukuran jarak variabel nominal antara $x_i$dan x_j ditunjukkan pada Persamaan (2.1),

$$\text{sim}(x_i, x_j) = \frac{1}{m}\sum_{l=1}^{m} S_{ijl} \qquad (2.1)$$

Dimana$S_{ijl} = 1$ jika $x_{il} = x_{jl}$ dan $S_{ijl} = 0$ jika $x_{il} \neq x_{jl}$

3. Metode pengukuran jarak untuk variabel kategorik ordinal

   Pada pengamatan dengan variabel ordinal maka pengukuran memiliki konsep yang digunakan sama dengan metode untuk data numerik, dimana kategorinya dinyatakan sebagai suatu bilangan bulat. Salah satu metode yang dapat digunakan untuk variabel ordinal adalah jarak *manhattan.* Dengan jumlah variable sebanyak *m,* maka rumus untuk pengukuran jarak $x_i$dan x_j pada variabel nominal ditunjukkan pada persamaan

   $$\text{sim}(x_i, x_j) = \frac{1}{m}\sum_{l=1}^{m}\left|x_{il} - x_{ij}\right| \qquad (2.2)$$

4. Metode pengukuran jarak untuk variabel numerik

   Pada variabel yang memiliki jenis data numerik maka jarak yang dapat digunakan adalah jarak *euclidean*. Misalkan terdapat dua observasi dengan variabel-variabel berdimensi m yaitu $x_i = x_1, x_2, \dots, x_m$ dan $x_j = x_1, x_2, \dots, x_m$

   Konsep jarak *euclidean* yang mengukur jarak antara observasi dan *x j* adalah sebagai berikut,

   $$d_{ij} = \sqrt{(x_i - x_j)^T (x_i - x_j)} \qquad (2.3)$$

   Terkait dengan pengertian dan tujuan dilakukannya analisis kelompok, dapat dinyatakan bahwa suatu kelompok (kelompok) yang baik adalah kelompok yang mempunyai. (Hair, Black, Babin, & Anderson, 2009).

   a. Homogenitas (kesamaan) yang tinggi antara anggota dalam satu kelompok (*within*-kelompok),
   b. Heterogenitas (perbedaan) yang tinggi antara kelompok yang satu dengan kelompok yang lain (*between* kelompok)

B. Metode Fuzzy C-Means

Fuzzy C-Means (FCM) adalah teknik pengelompokkan data yang memungkinkan suatu data menjadi dua atau lebih kelompok. Fuzzy C-means cluster pertama kali dikemukakan oleh Dunn (1973) dan kemudian dikembangkan oleh Bezdek (1981) yang banyak digunakan dalam *pattern recognition*. Metode ini merupakan pengembangan dari metode non hierarki K-means Cluster, karena pada awalnya ditentukan dulu jumlah kelompok atau cluster yang akan dibentuk. Kemudian dilakukan iterasi sampai mendapatkan keanggotaan kelompok tersebut. Metode ini adalah metode yang paling digemari karena merupakan metode yang paling robust ((Klawonn dan Höppner, 2001) dan (Klawonn, 2000)) dan memberikan hasil yang smooth (halus) dengan toleransi relatif (Shihab, 2000).

Suatu titik bisa memiliki keanggotaan parsial lebih banyak dari satu kelas Tidak boleh ada kelas kosong dan tidak ada kelas yang tidak berisi titik data FCM berusaha untuk menemukan karakteristik titik terbaik pada setiap kelompok, yang dapat dipertimbangkan sebagai "pusat" kelompok. Algoritma ini bekerja dengan menugaskan keanggotaan ke setiap titik data yang sesuai setiap pusat kelompok berdasarkan jarak antar pusat kelompok dan titik data. Algoritma *Fuzzy C-Means* didasarkan pada minimisasi fungsi objektif berikut (Velmurugan & Santhanam, 2010) adalah

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} \left(\mu_{ij}\right)^{m} \left\| x_i - v_j \right\|^2 \qquad (2.4)$$

Dimana $\left\| x_i - v_j \right\|$ adalah jarak Euclidean antara data ke i^{th} dan pusat kelompok j^{th}. Partisi *Fuzzy* dilakukan melalui iterasi optimalisasi fungsi objektif yang ditunjukkan di atas,

dengan memperbarui keanggotaan μ_{ij}dan pusat kelompok $v_j$oleh:

$$\mu_{ij} = 1/\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{ik}}\right)^{(2/m-1)} \quad (2.5)$$

$$v_j = \left(\sum_{i=1}^{n}\left(\mu_{ij}\right)^{m} x_i\right)/\left(\sum_{i=1}^{m}\left(\mu_{ij}\right)^{m}\right), \forall j = 1,2,\dots,c \quad (2.6)$$

Dimana $\{\mu_1,\dots,\mu_n\}$ adalah partisi *Fuzzy* c dan $\{v_1,\dots,v_n\}$ adalah himpunan *sentroid* 'm' adalah sebuah bilangan riil yang lebih dari 1 yang menyatakan derajat kekaburan (*degree of Fuzzyness*), indeks m ∈ [1, ∞], 'c' mewakili jumlah kelompok pusat, v_j mewakili pusat kelompok, μ_{ij} mewakili keanggotaan kelompok i^{th} ke kelompok j^{th}, d_{ij} mewakili jarak Euclidean antara data i^{th} dan pusat kelompok j^{th}.

C. Metode Fuzzy C-Modes

Pada sebagian besar algoritma pengelompokan *Fuzzy*, yang ditugaskan anggota data ke kelompok tidak jelas, tapi sentroid itu sendiri tidak *Fuzzy*. terdapat *Fuzzy K-Modes* Algoritma dimodifikasi sebagai algoritma *Fuzzy C-Modes* yang digunakan sentroid *Fuzzy* untuk mengelompokkan data kategorik. *Fuzzy sentroid* adalah seperangkat nilai *Fuzzy* yang mengandung nilai kategori setiap atribut. Untuk mengelompokkan data kategorik, diusulkan Algoritma *Fuzzy C-Modes* Algoritma memperluas algoritma *K-Modes* berdasarkan Prosedur tipe *C-Means Fuzzy*. Algoritma ini memperbarui pusat kelompok pada setiap iterasi dengan mengukur jarak antar masing-masing sentroid kelompok sentroid dan masing-masing objek. Misalkan $X = \{X_1, X_2, \dots, X_n\}$ menjadi himpunan n objek. Objek X_i diwakili sebagai $[x_{i1}, x_{i2}, \dots, x_{im}]$ dan $X_i = X_k$ jika $x_{ij} = x_{kj}$, $1 \le j \le m$. Algoritma *Fuzzy C-Modes* mengelompokkan data X ke dalam k kelompok dengan meminimalkan fungsi objektif

$$F(W,Z) = \sum_{l=1}^{k}\sum_{i=1}^{n} W_h^{\alpha} d(Z_1, X_i) \quad (2.7)$$

$$0 \le W_h \le 1; 1 \le l \le k; 1 \le i \le n,$$

$$\sum_{l=1}^{k} 1, 1 \le i \le n, \text{dan } 0 < \sum_{i=1}^{n} W_h < n, 1 \le l \le k$$

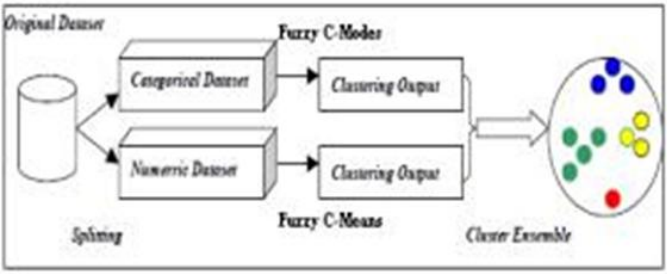
Sedangkan $*_{li}$ adalah derajat keanggotaan data X_i ke l^{th} kelompok, dan merupakan elemen matriks partisi $(k \times n)$. $W = \{*_{li}\} \cdot C^* = [C_1^*, C_2^*, \dots, C_l^*, \dots, C_k^*]$ dan C_l^* adalah pusat kelompok l^{th} dan parameter α mengontrol kekaburan dari tiap anggota objek. Algoritma *Fuzzy C-Modes* menggunakan ukuran ketidakmiripan yang sederhana untuk objek yang kategorik, anggap *X* dan *Y* adalah dua objek yang diwakili oleh $[x_1, x_2, \dots, x_m]$ dan $[y_1, y_2, \dots, y_m]$ maka $d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$ dimana $\delta(x_j, y_j) = \begin{cases} 0, x_j = y_j \\ 1, x_j \neq y_j \end{cases}$. Ukuran memenuhi ruang metrik objek-objek kategorik.

D. Pengelompokan Ensembel

Pengelompokan ensembel merupakan metode yang menggabungkan beberapa algoritma yang berbeda untuk mendapatkan partisi umum dari data, yang bertujuan untuk konsolidasi hasil pengelompokan individu (Suguna & Selvi, 2012). Tujuan pengelompokan ensembel adalah untuk menggabungkan hasil pengelompokan dari beberapa algoritma pengelompokan untuk mendapatkan hasil pengelompokan yang lebih baik dan *robust* (Yoon, Ahn, Lee, Cho, & Kim, 2006). Pengelompokan ensembel terdiri atas dua tahap algoritma. Tahap pertama adalah melakukan pengelompokan dengan beberapa algoritma dan menyimpan hasil pengelompokan tersebut. Kedua, menggunakan fungsi konsensus untuk menentukan *final* kelompok dari kelompok-kelompok hasil tahap pertama. Langkah-langkah

dalam analisis data campuran menggunakan metode pengelompokan ensembel yang disebut Algoritma CEBMDC memiliki tahapan sebagai berikut, (He, Xu, dan Deng, 2005b)

1. Membagi data menjadi dua subdata, yaitu murni numerik dan murni kategorik.
2. Melakukan pengelompokan objek dengan variabel numerik dengan algoritma pengelompokan data numerik, serta melakukan pengelompokan objek dengan variabel kategorik dengan algoritma pengelompokan data kategorik.
3. Menggabungkan (*combining*) hasil pengelompokan dari variabel numerik dan kategorik, yang disebut proses ensembel.
4. Melakukan pengelompokan ensembel menggunakan algoritma pengelompokan data kategorik untuk mendapatkan kelompok akhir (*final* kelompok).



Gambar 1 *Overview* Pengelompokan Ensembel

Metode ensembel *Fuzzy* yang digunakan dengan langkah-langkah berikut ini:

1. Langkah-langkah algoritma *Fuzzy* C-Means
   a. Menetapkan nilai dari tiap μ_{ij} , dimana $\mu_{ij} \geq 0$
   b. Melakukan iterasi pada Langkah 1
   c. Menghitung sentroid kelompok v_j menggunakan Persamaan 2.6

d. Menggunakan v_j yang terbaru, perbarui nilai μ_{ij} oleh Persamaan 2.5

2. Langkah-langkah algoritma *Fuzzy C-Modes*
   a. Secara acak menetapkan label kelompok untuk setiap objek, yaitu menginisialisasi keanggotaan kelompok $W^{(1)}$ Tentukan $C^{*(1)}$ sehingga meminimalkan $F\left(W^{(1)}, C^{*(1)}\right)$ jika $t = 1$
   b. Menentukan $W^{(t+1)}$ sedemikian hingga $F\left(W^{(t+1)}, C^{*(t)}\right)$ paling minimal. Jika $F\left(W^{(t+1)}, C^{*(t)}\right) = F\left(W^{(t)}, C^{*(t)}\right)$. Jika tidak $t = t + 1$ dan menuju pada Langkah 3
   c. Menentukan $C^{*(t+1)}$ sehingga meminimalkan $F\left(W^{(t+1)}, C^{*(t+1)}\right)$. Jika $F\left(W^{(t+1)}, C^{*(t+1)}\right) = F\left(W^{(t+1)}, C^{*(t)}\right)$ maka berhenti, jika tidak kembali pada Langkah 2

E. Metode Kprototypes

K-Prototype adalah metode pengelompokan pada data dengan tipe campuran numerik dan kategorikal. Algoritma ini dipilih karena sangat sederhana dari sisi kompleksitas algoritma dan mampu menangani data dengan ukuran yang sangat besar. *K-Prototype* adalah salah satu metode Pengelompokan yang berbasis partitioning. Algoritma ini adalah hasil pengembangan atau kombinasi antara *K-Means* dan *K-Modes* Pengelompokan untuk menangani Pengelompokan pada data dengan campuran atribut bertipe numerik dan kategorik (Huang, 1997). Perubahan yang mendasar terdapat pada pengukuran kesamaan (*similarity measure*) antara *object* dengan *sentroid*

(prototype)-nya. Pada proses pengelompokan dengan *K-Prototype* disini dilakukan beberapa proses utama yaitu :

1. Inisialiasi awal prototype

Pada proses ini akan dilakukan pemilihan sejumlah *K- Prototype* dari dataset X sesuai dengan jumlah kelompok yang ditentukan. Pemilihan ini biasanya dilakukan secara acak. Proses ini akan menghasilkan matrik *prototype* berukuran k kali panjang atribut X.

2. Alokasi object di dalam X ke Kelompok dengan prototype terdekat

Pada proses ini akan dilakukan pengalokasian semua object di dalam dataset X ke kelompok yang memiliki jarak *prototype* terdekat dengan object yang diukur. Untuk setiap kali object X selesai dialokasikan, maka selanjutnya akan dilakukan penghitungan (update) terhadap prototype kelompok yang berkaitan.

3. Realokasi object jika terjadi perubahan prototype

Setelah semua object dalam X selesai dialokasikan, selanjutnya akan dilakukan pengukuran ulang jarak antara semua object di dalam X terhadap semua prototype yang ada. Jika ditemukan adanya object yang ternyata lebih dekat ke prototype yang lain, maka akan dilakukan pemindahan keanggotaan dan kemudian akan dilakukan update terhadap prototype kelompok lama dan prototype kelompok baru. Proses ini akan terus dilakukan sampai tidak ada lagi perubahan prototype. Misalkan $X = \{X_1 , X_2 , \ldots , X_n\}$ adalah kumpulan n objek dan $X_i = [X_{i1} , X_{i2} , \ldots , X_{im}]^T$, dimana m menunjukan atribut dan I menunjukkan kelompok ke-i.

Sebelum masuk proses algoritma *K-Prototypes* tentukan jumlah *k* yang akan dibentuk batasannya minimal 2 dan maksimal $\sqrt{n}$atau $n/2$ dimana *n* adalah jumlah data poin atau obyek. (Lin et al, 2005)

- Tahap 1: Menentukan *k* dengan inisial kluster $z_1, z_2, \dots, z_k$ secara acak dari n buah titik $\{x_1, x_2, \dots, x_n\}$
- Tahap 2: Menghitung jarak seluruh data *point* pada dataset terhadap inisial kluster awal, alokasikan data point ke dalam kelompok yang memiliki jarak *prototype* terdekat dengan objek yang diukur.
- Tahap 3: Menghitung titik pusat kelompok yang baru setelah semua objek dialokasikan. Lalu realokasikan semua titik data pada dataset terhadap *prototype* yang baru
- Tahap 4: Jika titik pusat kelompok tidak berubah atau sudah konvergen maka proses algoritma berhenti tetapi jika titik pusat masih berubah-ubah secara signifikan maka proses kembali ke tahap 2 dan 3 hingga iterasi maksimum tercapai atau sudah tidak ada perpindahan objek.

F. Ukuran Kesamaan (Similarity Measure)

Bentuk umum ukuran kesamaan dinyatakan sebagai berikut

$$d\left(X_j, Z_i\right) = \sum_{j=1}^{m} \delta(x_{jl}, z_{il}) \tag{2.8}$$

$z_i = \left[z_{i1}, z_{i2}, \dots, z_{im}\right]^T$ adalah *prototype* untuk kelompok *i*.

Ukuran kesamaan untuk atribut numerik dikenal dengan jarak euclidean ditunjukkan dalam persamaan (2.9) berikut ini

$$d\left(X_j, Z_i\right) = \left(\sum_{l=1}^{m_r} ({x_{jl}}^r - {z_{il}}^r)^2 \right)^{1/2} \quad (2.9)$$

x_{jl}^r adalah jumlah atribut numerik l, z_{il}^r adalah rata-rat atau *prototype* atribut numerik ke *l* kelompok *i*.m_r adalah jumlah atribut numerik. Sedangkan ukuran kesamaan untuk data kategorikal adalah

$$d\left(X_j, Z_i\right) = \gamma_i \sum_{l=l+1}^{m_c} \delta({x_{jl}}^c, {z_{il}}^c) \quad (2.10)$$

Dan *simple matching similarity measure* untuk data kategorikal adalah

$$\delta({x_{jl}}^c, {z_{il}}^c) = \begin{cases} 0 & \left({x_{jl}}^c - {z_{il}}^c\right) \\ 1 & \left({x_{jl}}^c \neq {z_{il}}^c\right) \end{cases} \quad (2.11)$$

$\delta(x_{ij}^c, z_{il}^c)$ adalah bobot untuk atribut kategori pada kelompok *i* yang nilainya merupakan nilai standar deviasi untuk atribut numerik pada masing-masing kelompok. Ketika ${x^c}_{ij}$ adalah nilai atribut kategorikal z_{il}^c adalah modus atribut ke *l* kelompok *i*. m^c adalah jumlah atribut kategorikal. He memodifikasi *simple matching similarity measure* menjadi persamaan (5) untuk meningkatkan kemiripan objek dalam kelompok dengan atribut kategorikal sehingga hasil pengelompokan menjadi lebih baik (Amir & Lipika, 2007).

$$\delta({x_{jl}}^c, {z_{il}}^c) = \begin{cases} 1 - x_{jl}^c & \left({x_{jl}}^c - {z_{il}}^c\right) \\ 1 & \left({x_{jl}}^c \neq {z_{il}}^c\right) \end{cases} \quad (2.12)$$

$\omega(x_{jl}^c, i)$ adalah nilai penimbang untuk x_{jl}^c dimana nilai $\omega(x_{jl}^c, i)$ adalah