Ensembel *Fuzzy*, Ensembel Rock pada Pengelompokan Pelamar Bidikmisi Se-Jawa Timur Tahun 2017

Laila Qadrini
Universitas Sulawesi Barat, Majene, laila.qadrini@unsulbar.ac.id

Fardinah
Universitas Sulawesi Barat, Majene, fardinah@unsulbar.ac.id

Meryta FF
Universitas Sulawesi Barat, Majene, merytaff@unsulbar.ac.id

ABSTRAK, 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 atau sebaliknya. Selain pengelompokan dengan metode transformasi tersebut, dikembangkan sebuah metode pengelompokan ensembel untuk data campuran. Pengelompokan ensembel adalah teknik pengelompokan untuk menggabungkan hasil pengelompokan beberapa algoritma pengelompokan dengan tujuan untuk mendapatkan hasil kelompok yang lebih baik, berdasarkan indeks validitas internal kelompok yaitu nilai SSW, Rata-rata koefisien *Silhouette* dan nilai Indeks Dunn yang dianalisis untuk 2,3 dan 4 Kelompok, diperoleh bahwa metode Ensembel *Fuzzy* lebih baik dan tepat digunakan pada data campuran yang ada pada penelitian ini daripada metode pengelompokan Ensembel ROCK.

Kata Kunci: *Pengelompokan, Ensembel Fuzzy, Ensembel ROCK*

1. PENDAHULUAN

Metode pengelompokan dalam data mining berbeda dengan metode konvensional yang biasa digunakan untuk pengelompokan. 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 et.al, 2010) melakukan transformasi variabel kategorik ke dalam bentuk numerik, kemudian pengelompokan objek dilakukan dengan metode pengelompokan data numerik, selain pengelompokan dengan metode transformasi tersebut, dikembangkan sebuah metode pengelompokan ensembel untuk data campuran. (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. Hasil tersebut menunjukkan bahwa metode ensembel ROCK memberikan hasil pengelompokan lebih baik daripada metode ensembel SWFM, J. (Suguna dan 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*). dievaluasi dengan ukuran f-*measure* dan *entropy*. Namun, penelitian-penelitian tersebut tidak dilakukan perbandingan hasil pengelompokan antara kedua metode. Oleh karena itu, pada penelitian ini dilakukan perbandingan hasil pengelompokan antara metode pengelompokan ensembel *Fuzzy*, dan ensembel Rock Hasil pengelompokan kedua metode tersebut dapat dilihat berdasarkan ukuran validitas kelompok. (Steinbach et.al, 2000).

2. TINJAUAN PUSTAKA

Metode Fuzzy C-Means

Algoritma *Fuzzy C-Means* didasarkan pada minimisasi fungsi objektif berikut (Velmurugan dan Santhanam, 2010) adalah

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

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:

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

$$v_j=\left(\sum_{i=1}^{n}(\mu_{ij})^m x_i\right)/\left(\sum_{i=1}^{n}(\mu_{ij})^m\right),\forall j=1,2,...,c \qquad (3)$$

dimana $\{\mu_1,...,\mu_n\}$ adalah partisi *Fuzzy c* dan $\{v_1,...,v_n\}$ 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}.

Metode Fuzzy C-Modes

Algoritma ini memperbarui pusat kelompok pada setiap iterasi dengan mengukur jarak antar masing-masing sentroid kelompok sentroid dan masing-masing objek. Misalkan $X=\{X_1,X_2,...,X_n\}$ menjadi impunan *n* objek. Objek X_i diwakili sebagai $[x_{i1},x_{i2},...,x_{im}]$ dan $X_i=X_k$ jika $x_{i,j}=x_{k,j},1\le j\le m.$ Algoritma *Fuzzy C-Modes* mengelompokkan data *X* ke dalam *k* kelompok dengan meminimalkan fungsi objektif

$$F(W,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n}W_h^{\alpha}d\left(Z_1,X_i\right) \qquad (4)$$

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

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

Sedangkan $*_{li}$ adalah derajat keanggotaan data X_i ke l^{th} kelompok, dan merupakan elemen matriks partisi $(k\times n)$. $W=[*_{li}].C*=[C*_1,C*_2,...,C*_l,....,C*_k]$ dan $C*_l$ adalah pusat kelompok l^{th} dan parameter α mengontrol kekaburan dari tiap anggota objek.

Metode DBSCAN

Algoritma DBSCAN

Himpunan_Klaster = DBSCAN($\varepsilon, MinObj$)

1. Tandai semua objek sebagai *unvisited*
   Repeat
2. Pilih secara acak sebuah objek p dari semua objek yang berlabel *unvisited*
3. Tandai p sebagai *visited*

If dalam radius ε objek p memiliki minimal *MinObj* objek
then Buat sebuah klaster baru C Tambahkan p kedalam C Masukkan semua objek yang menjadi tetangga p ke dalam N
for setiap objek p' di N
do if p' berlabel *unvisited*
then tandai p' sebagai *visited*
if dalam radius ε p' memiliki minimal *MinObj* objek
then tambahkan semua objek dalam radius ε tersebut ke dalam N.
if p' bukan anggota dari klaster manapun
then tambahkan p' kedalam C
end Keluarkan C sebagai sebuah klaster *output*
Else tandai p sebagai derau
Until tidak ada objek yang berlabel *unvisited*

Metode ROCK

Pengelompokan data kategorik dengan algoritma ROCK dilakukan dengan tiga langkah. Langkah pertama adalah menghitung similaritas. Ukuran kemiripan antara pasangan objek ke-*i* dan objek ke-*j* dihitung dengan rumusan yang didefinisikan pada persamaan 6.

$$sim(X_i,X_j)=\frac{|X_i\cap X_j|}{|X_i\cup X_j|},\ i\ne j \qquad (6)$$

$i=1,2,3,\dots,n$ dan $j=1,2,3,\dots,n$

X_i: Himpunan pengamatan ke-i dengan

$$X_i = \{x_{1i}, x_{2i}, x_{3i}, \dots, x_{m_{kategorik}i}\}$$

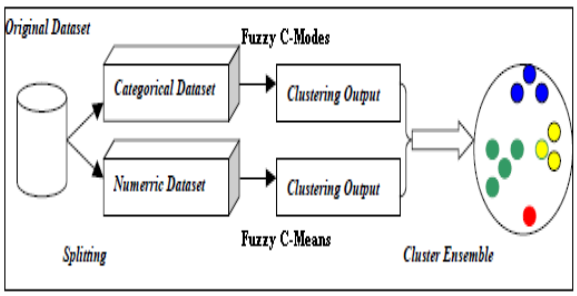
X_j: Himpunan pengamatan ke-j dengan

$$X_j = \{x_{1j}, x_{2j}, x_{3j}, \dots, x_{m_{kategorik}j}\}$$

$|X|$: Bilangan cardinal atau jumlah anggota dari himpunan X. Langkah kedua adalah menentukan tetangga, pengamatan dinyatakan sebagai tetangga jika nilai $sim(X_i, X_j) \geq \theta$. Langkah terakhir adalah menghitung *link* antar objek pengamatan. Besarnya *link* dipengaruhi oleh nilai *threshold* (θ) yang merupakan parameter yang ditentukan oleh pengguna yang dapat digunakan untuk mengontrol seberapa dekat hubungan antara objek. Besarnya nilai θ yang diinputkan adalah $0 < \theta < 1$. Algoritma ROCK berhenti ketika jumlah dari kelompok yang diharapkan sudah terpenuhi atau tidak ada lagi *link* antara kelompok-kelompok. (Dutta, 2005)

Pengelompokan Ensembel

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.



Gambar 1. *Overview* Pengelompokan Ensembel

Uji Validitas kelompok

1. Kelompok *cohesion* mengukur seberapa dekat obyek-obyek yang berada dalam satu kelompok,

$$WSS = \sum_k \sum_{x \in C_k} (x - \bar{x}_k)^2 \qquad (7)$$

Kelompok *separation* digunakan untuk mengukur seberapa berbeda kelompok-kelompok yang terbentuk.

$$BSS = \sum_k |C_k| (x - \bar{x}_k)^2 \qquad (8)$$

Dimana $|C_k|$adalah kelompok ke-k, dan $\bar{x}$ adalah rata-rata objek pengamatan.

2. *Indeks Dunn* (D)

$$D = \min_{1 \leq l \leq n} \left\{ \min_{1 \leq j \leq n, i \neq j} \left\{ \frac{d(c_i, c_j)}{\max_{1 \leq k \leq n} (d'(c_k))} \right\} \right\} \qquad (9)$$

dengan $d(c_i, c_j)$ = jarak antar kelompok c_i dan c_j, $d'(c_k)$= jarak dalam kelompok c_k. Nilai terbesar dari D diambil sebagai jumlah optimum kelompok (Azuaje dan Bolshakova, 2001).

3. *Indeks Global Silhouette*

Untuk mendapatkan indeks kelompok $S(i)$ digunakan rumus berikut:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \qquad (10)$$

dengan $a(i)$= rata-rata perbedaan dari i-obyek dengan semua obyek lain pada kelompok yang sama. $b(i)$= obyek pada kelompok lain (di kelompok terdekat).

3. METODOLOGI

Menyusun Algoritma Metode Ensembel *Fuzzy*

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

Langkah-langkah algoritma* Fuzzy *C-Means

1. Menetapkan nilai dari tiap $\mu_{ij,}$ dimana $\mu_{ij,} \geq 0$
2. Melakukan iterasi pada Langkah 1
3. Menghitung sentroid kelompok v_j menggunakan Persamaan 3
4. Menggunakan v_j yang terbaru, perbarui nilai μ_{ij} oleh Persamaan 2

***Langkah-langkah algoritma* Fuzzy C-Modes**

1. Secara acak menetapkan label kelompok untuk setiap objek, yaitu menginisialisasi keanggotaan kelompok $W^{(1)}$ Tentukan $C^{*(1)}$ sehingga meminimalkan $F(W^{(1)}, C^{*(1)})$ jika t =1

2. Menentukan $W^{(t+1)}$ sedemikian hingga $F\left(W^{(t+1)}, C^{*(t)}\right)$ paling minimal. Jika $F\left(W^{(t+1)}, C^{*(t)}\right) = F\left(W^{(t)}, C^{*(t)}\right)$ maka berhenti. Jika tidak $t = t+1$ dan menuju pada Langkah 3
3. 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.

4. PEMBAHASAN

Tabel 1. Perbandingan Hasil Analisis Ensembel *Fuzzy* dan Ensembel ROCK untuk 2 Kelompok

Metode	Indeks Validasi Kelompok	Nilai
Ensembel *Fuzzy*	SSW	123.476
	Rata-rata *Silhouette*	0,668
	Dunn	0,008
Ensembel ROCK	SSW	110.278
	Rata-rata *Silhouette*	0,480
	Dunn	0,003

Tabel 2. Perbandingan Hasil Analisis Ensembel *Fuzzy* dan Ensembel ROCK untuk 3 Kelompok

Metode	Indeks Validasi Kelompok	Nilai
Ensembel *Fuzzy*	SSW	692.371
	Rata-rata *Silhouette*	0,604
	Dunn	0,003
Ensembel ROCK	SSW	429.21
	Rata-rata *Silhouette*	0,231
	Dunn	0,092

Tabel 3. Perbandingan Hasil Analisis Ensembel *Fuzzy* dan Ensembel ROCK untuk 4 Kelompok

Metode	Indeks Validasi Kelompok	Nilai
Ensembel *Fuzzy*	SSW	420.111
	Rata-rata *Silhouette*	0,578
	Dunn	0,009
Ensembel ROCK	SSW	230.435
	Rata-rata *Silhouette*	0,327
	Dunn	0,0006

Tabel diatas dapat menjelaskan bahwa berdasarkan indeks validitas internal kelompok yaitu nilai SSW, Rata-rata koefisien *Silhouette* dan nilai Indeks Dunn, diperoleh bahwa metode Ensembel *Fuzzy* lebih baik dan tepat digunakan pada data campuran yang ada pada penelitian ini daripada metode pengelompokan Ensembel ROCK.

5. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan, maka diperoleh beberapa kesimpulan antara lain,

1. Algoritma yang dibangun untuk melakukan analisis agar dapat menangani permasalahan dalam melakukan pengelompokan data berskala campuran numerik dan kategorik. Pemrograman pada *software R Project* yang dibangun dapat mempermudah dilakukannya analisis dan perbandingan dari kedua metode tersebut.
2. Berdasarkan dengan perbandingan indeks validasi pengelompokan dapat dikatakan bahwa metode ensembel *fuzzy* lebih tepat digunakan pada data penelitian ini dibandingkan metode Ensembel Rock

6. DAFTAR PUSTAKA

[1] Alvionita, (2017), *Metode Ensembel Rock dan SWFM untuk* pengelompokan *data campuran kategorik dan numerik pada kasus aksesi jeruk,* Tesis Program Magister FMIPA, Statistika, Institut Teknologi Sepuluh Nopember, Surabaya.

[2] Azuaje, F., dan Nadia, B., (2001), "Improving Expression Data Mining through Cluster Validity", Departement of Computer Science. Trinity College Dublin. Irlandia.

[3] Bolshakova, N., (2003), "Cluster Validity Algorithms", Departement of Computer Science.TrinityCollege Dublin Irlandia.

[4] Dewangan, R. R., Sharma, L. K., dan Akasapu, A. K., (2010), "*Fuzzy Clustering Technique for Numerical and Categorical Dataset*", *International Journal on Computer Science and Engineering*, hal 75-80.

[5] Guha, S., Rastogi, R., dan Shim, K., (2000), "ROCK: A Robust Clustering Algorithm for Categorical Attributes", *Proceedings of the 15th International Conference on Data Engineering.*

[6] He, Z., Xu, X., dan Deng, S., (2005a), "A Cluster Ensemble Method For Clustering Categorical Data", *Information Fusion, 6*, hal 143-151.

[7] He, Z., Xu, X., dan Deng, S., (2005a), "A Kelompok Ensemble Method For Pengelompokan Categorical Data", *Information Fusion, 6*, hal 143-151.

[8] Huang, Z.X, "Pengelompokan Large Data Sets with Mixed and Numeric and Categorical values", *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '97)*, hal. 21–34, 1997.

[9] Johnson, R. A., dan Winchern, D. W., (2007), *Applied Multivariate Statistical Analysis* (sixth ed.), Pearson Education, Inc, New Jersey.

[10] Suguna, J., dan Selvi, M. A., (2012), "Ensemble *Fuzzy* Pengelompokan for Mixed Numerical and Categorical Data", *International Journal of Computer Application*, vol 42, no 3.

[11] Suyanto, (2017), *Data Mining untuk Klasifikasi dan Klasterisasi data*, Informatika Bandung, Bandung.

[12] Velmurugan, T., dan Santhanam, T., (2010), "Clustering Mixed Data Points using *Fuzzy* C-Means Algorithm for Performance Analysis", *International Journal on Computer Science and Engineering*, vol. 2, no. 9.