K-MEANS DAN *FUZZY C-MEANS* PADA ANALISIS DATA POLUSI UDARA DI KOTA X

Sandi Fajar Rodiyansyah

Teknik Informatika Universitas Majalengka
Gd. Fakultas Teknik UNMA Jl. KH. Abdul Halim No. 103 Majalengka 45418 – Jawa Barat
Email : sfr@ft.unma.ac.id

Abstrak

Polusi udara adalah masalah umum bagi daerah perkotaan. Dimana umumnya daerah perkotaan memiliki tingkat produksi gas polutan lebih besar dibanding daerah lainnya. Pada penelitian ini, dilakukan kaji banding algoritma k-means dan fuzzy c-means pada analisis cluster data polusi udara harian di suatu kota. Hasil pengujian menunjukan bahwa rata-rata standar deviasi pada hasil clustering fuzzy c-means lebih kecil daripada rata-rata standar deviasi pada hasil clustering k-means. Selain itu, penelitian ini juga menghasilkan simpulan bahwa bahwa parameter natrium dioksida (NO2), non metal hydro carbon (NMHC) dan natrium oksida (NOx) memiliki pengaruh yang signifikan terhadap proses clustering.

Kata kunci:*polusi udara, clustering, k-means, fuzzy c-means.*

1. Pendahuluan

Indonesia memiliki kepadatan penduduk yang melampaui angka 12,4% pada tahun 1950 hingga 48,1% pada tahun 2005. Angka ini diperkirakan akan mencapai 58,5% pada tahun 2050, hanya mencakup kawasan perkotaan itu sendiri [1]. Pembangunan secara masif pada sebuah kota telah berdampak pada kehidupan manusia penghuni kota tersebut, baik secara lahir maupun batin. Kawasan perkotaan menjadi salah satu identitas keruangan pada negara berkembang, tidak hanya yang berkenaan dengan inovasi modern namun juga yang berhubungan dengan permasalah lingkungan dan ekologis [2]. Oleh karena itu, satu indikator dari kemajuan sebuah kota adalah kondisi lingkungannya.

Sementara itu, kondisi lingkungan khususnya kondisi udara sangat dipengaruhi oleh faktor-faktor, diantaranya adalah faktor transportasi, faktor industri dan faktor pembangunan infrastruktur kota tersebut. Udara bersih adalah salah satu kebutuhan dasar manusia untuk dapat melanjutkan hidup dan kehidupannya. Dengan udara yang bersih, manusia dapat beraktifitas dengan baik.

Setiap kota memiliki pola masing-masing pada kondisi udara di kota tersebut. Hal ini dipengaruhi oleh pola aktifitas masyarakatnya, khususnya pola aktifitas transportasi. Sehingga, setiap jam kondisi udara akan berbeda dengan jam yang lain, bahkan di hari yang lain. Mengingat data kondisi udara akan berjumlah banyak terutama jika pengambilan data dilakukan setiap jam maka diperlukan teknik untuk mengenali pola keadaan udara di suatu kota dengan bantuan komputer.

Berdasarkan masalah tersebut, maka penelitian ini akan mencoba melakukan analisis perbandingan algoritma *k-mean* dan *fuzzy c-mean* untuk digunakan pada proses analisis cluster menggunakan data polusi udara harian di suatu kota. Data ini adalah data dari hasil penelitian De Vito [3] yang diambil disalah satu kota di Italia.

Sebelumnya telah banyak dilakukan penelitian *clustering* data pencemaran udara. Salah satunya adalah yang dilakukan oleh Yanti dan Ulfah [4] melakukan pengelompokan pulutan berdasarkan beban polutan yang mengandung zat-zat kimia berbahaya yang dihasilkannya. Pengelompokan polutan ini menggunakan jaringan syaraf tiruan dengan metode *learning vector quantization* (LVQ) menghasilkan *learning rate* 0.0011719 dengan target *error* 0.001 tercapai pada epoch ke-10. Sementara itu, Rachmatin [5] melakukan analisis metode-metode *agglomerative* diantaranya *single linkage method, complete linkage method, average linkage method. Ward's method, centroid method* dan *median method* yang diterapkan pada data tingkat polusi udara yang meghasilkan bahwa masing-masing metode tersebut memberikan jumlah cluster yang berbeda. Sitepu dkk. [6] melakukan penelitian yang membahas pengelompokan 10 jenis industri yang berada di Sumatera Selatan berdasarkan jenis polutan yang dihasilkan dan mengetahui ciri-ciri dari setiap kelompok industri. Berdasarkan hasil analisis cluster hierarki ada 3 kelompok industri yaitu : cluster pertama: industri karet, sawit, pengalengan ikan, listrik, pertambangan dan semen. Cluster kedua : industri migas, minyak goreng dan makanan. Cluster ketiga : industri pupuk. Pada metode non-hierarki, cluster pertama yaitu industri yang memiliki rata-rata polutan lebih besar daripada cluster kedua. Anggotanya adalah industri migas, minyak goreng, makanan dan pupuk. Cluster kedua yaitu industri yang memiliki rata-rata polutan lebih kecil daripada cluster pertama. Anggotanya adalah industri karet, sawit, pengalengan ikan, listrik, pertambangan dan semen.

2. Landasan Teori

Data mining merupakan suatu metode untuk menemukan pengetahuan dalam suatu tumpukan data yang cukup besar. Data mining adalah proses menggali dan menganalisa sejumlah data yang sangat besar untuk

memperoleh sesuatu yang benar, baru dan bermanfaat dan akhirnya dapat ditemukan suatu corak atau pola dalam data tersebut. Han dan Kamber [7]. Salah satu metode dalam data mining adalah *clustering*, yaitu pengelompokan data berdasarkan kemiripan atau kesamaannya. Metode ini diterapkan apabila data akan dibagi menjadi cluster yang terbentuk secara alami dan tidak ada prediksi klasifikasi [8].

K-means adalah salah satu dari beberapa algoritma *clustering* dengan menggunakan konsep dasar bahwa semakin dekat dengan pusat cluster maka data termasuk kedalam kategori cluster tersebut. Langkah-langkah dalam algoritma k-means adalah sebagai berikut :

1. Tentukan jumlah titik cluster (k) secara acak.
2. Hitung jarak masing-masing data dengan semua titik cluster dengan menggunakan persamaan (1).

$$d(x_i, x_j) = \sqrt{(x_{i1}\text{-}x_{j1})^2 + \cdots + (x_{in}\text{-}x_{jn})^2} \quad (1)$$

3. Menentukan anggota masing-masing cluster berdasarkan jarak terdekat
4. Menghitung ulang pusat cluster dengan menggunakan persamaan (2).

$$\mu^k = \frac{\sum_{q=1}^{n_k} x^q}{N^k} \quad (2)$$

Langkah 2 sampai langkah 4 diulangi sampai dengan pusat cluster tidak berubah [8]

Konsep dasar *fuzzy c-means*, pertama kali adalah menentukan pusat cluster, yang akan menandai lokasi rata-rata untuk tiap-tiap cluster. Pada kondisi awal, pusat cluster ini masih belum akurat. Tiap-tiap titik data memiliki derajat keanggotaan untuk tiap-tiap cluster [9]. Algoritma *fuzzy c-means* dapat dilihat pada Gambar 1.

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

3. Update $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. If $\| U^{(k+1)} - U^{(k)} \| < \xi$ then STOP; otherwise return to step 2.

Gambar 1 : *Algoritma fuzzy c-means*[7]

3. Metode Penelitian

Tahapan-tahapan yang dilakukan dalam penelitian ini adalah sebagai berikut :

1. Proses pengumpulan data.
   Data diperoleh dari hasil penelitian De Vito [3].
2. Pembersihan data dan transformasi ke *spreadsheet.*
   Data yang diperoleh dari tahap pertama masih banyak yang memiliki data yang kosong (*missing value*). Oleh karena itu, perlu dilakukan pemberihan data selanjutnya dilakukan transformasi menjadi *file spreadsheet.*
3. Proses *clustering* dengan *k-mean* dan *fuzzy c-means*
   Pengujian dilakukan dengan menggunakan *software* RapidMiner Versi 7.1
4. Penarikan kesempulan
   Dilakukan penarikan kesimpulan dari hasil *clustering* yang telah dilakukan pada tahapan 3. Penarikan kesimpulan dilakukan dengan melakukan pengujian hasil *clustering* meggunakan nilai rata-rata dari standar deviasi pada setiap parameter anggota masing-masing cluster. Standar deviasi digunakan karena nilai standar deviasi mencerminkan tingkat keseragaman data, semakin seragam nilai standar deviasi akan semakin kecil.

4. Pembahasan

Data yang digunakan pada proses *clustering* ini adalah data pengamatan hasil pengambilan dengan menggunakan sensor pada penelitian [3] di salah satu kota di Italia yang diambil pada periode Maret 2004 sampai dengan Februari 2005 yang mengambil parameter-parameter sebagai berikut :

1. Tanggal
2. Waktu
3. CO (carbon monoksida) dengan satuan mg/m^3
4. Non Metanic Hydro Carbons dengan satuan microg/m^3.
5. C6H6 (benzene) dengan satuan microg/m^3.
6. NOx (nitrogen oksida) dengan satuan ppb (part per billion).
7. NO2 (nitrogen dioksida) dengan satuan microg/m^3.
8. Temperatur dengan satuan derajat celcius.
9. Relative Humidity dengan satuan %.
10. Absolute Humidity dengan satuan g/m^3.

Data penelitian [3] menghasilkan data sebanyak 9358. Namun hasil penelitian tersebut masih banyak data yang bernilai kosong (*missing value*) sehingga sebelum dilakukan *clustering* diperlukan proses pembersihan data. Gambar 1 adalah sebagian data asli yang akan digunakan pada proses *clustering* pada penelitian ini.

```
1  Date;Time;CO(GT);PT08.S1(CO);NMHC(GT);C6H6(GT);PT08.S2(NMHC);NOx(GT);PT08.S3(NOx);NO2(GT
2  10/03/2004;18.00.00;2.6;1360;150;11.9;1046;166;1056;113;1692;1268;13.6;48.9;0.7578;;
3  10/03/2004;19.00.00;2;1292;112;9.4;955;103;1174;92;1559;972;13.3;47.7;0.7255;;
4  10/03/2004;20.00.00;2.2;1402;88;9.0;939;131;1140;114;1555;1074;11.9;54.0;0.7502;;
5  10/03/2004;21.00.00;2.2;1376;80;9.2;948;172;1092;122;1584;1203;11.0;60.0;0.7867;;
```

Gambar 2 : *Data penelitian*

Tahapan selanjutnya adalah dilakukan pembersihan data. Pembersihan data ini dilakukan dengan tujuan agar data yang diperloleh dari penelitian sebelumnya dibersihkan dari nilai-nilai yang hilang (*missing value*). Disamping melakukan pembersihan pada tahapan ini juga dilakukan proses konversi data. Proses ini dilakukan karena data asli yang diperoleh disimpan dengan format CSV

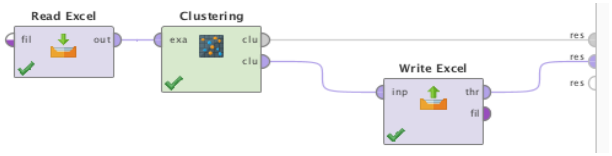
(*comma-separated values*) dan harus dikonversi menjadi format *spreadsheet* (.xlsx). Hal ini dilakukan karena program k-*means* dan *fuzzy* c-*means* yang terdapat pada *software* Rapid Miner 7.1 yang sudah dirancang sebelumnya didesain untuk membaca file data yang berekstensi .xlsx. Gambar 3 adalah sebagian dari data yang telah dilakukan pembersihan dan konversi menjadi *spreadsheet.*

	A	B	C	D	E	F	G	H	I	J
1	Date	Time	CO(GT)	NMHC(GT)	C6H6(GT)	NOx(GT)	NO2(GT)	T	RH	AH
2	10/03/2004	18.00.00	2.6	150	11.9	166	113	13.6	48.9	0.7578
3	10/03/2004	19.00.00	2	112	9.4	103	92	13.3	47.7	0.7255
4	10/03/2004	20.00.00	2.2	88	9	131	114	11.9	54	0.7502
5	10/03/2004	21.00.00	2.2	80	9.2	172	122	11	60	0.7867

Gambar 3 : *Data penelitian dalam spreadsheet*

1. Proses *clustering* menggunakan k-means

Proses *clustering* menggunakan k-means dilakukan dengan menggunakan *software* Rapid Miner 7.1. Proses ini dilakukan dengan menggunakan operator dengan konfigurasi yang terlihat pada gambar 4.



Gambar 4 : *Operator Rapid Miner k-means*

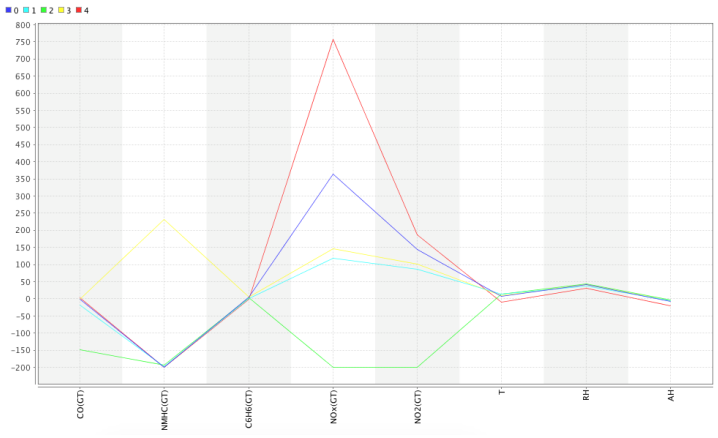
Opearator Read Excel digunakan untuk membaca *spreadsheet* yang berisi data yang sudah melewati tahap pembersihan data. Selanjutnya dilakukan *clustering* menggunakan k-*means* pada opeator Clustering. Setelah itu, hasil *clustering* disimpan pad *spreadsheet* dengan penggunakan operator Write Excel.

Proses *clustering* tersebut menggunakan parameter k=5 dan maxruns=100 iterasi sehingga menghasilkan nilai rata-rata setiap parameter untuk setiap cluster yang dapat dilihat pada tabel 1.

Tabel 1. *Pusat cluster pada k-means*

Par	**C0**	**C1**	**C2**	**C3**	**C4**
CO	-0.541	-18.60	-149.29	-1.177	3.938
NMHC	-200.0	-198.50	-194.28	232.64	-200.0
C6H6	4.340	-0.079	3.168	4.400	-0.129
NOx	363.731	119.48	-200.0	145.40	757.34
NO2	144.162	86.67	-200.0	101.15	187.16
T	7.008	12.929	14.029	9.025	-9.037
RH	40.759	37.683	44.64	41.57	31.855
AH	-7.258	-5.708	-4.232	-5.315	-19.80

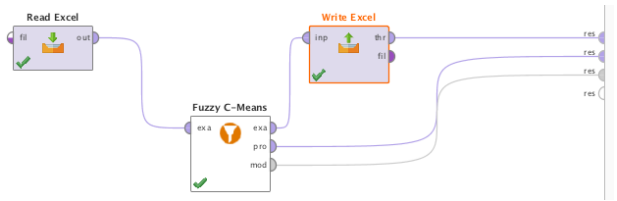
Dengan sebaran anggota cluster 0 sebanyak 2061 data, anggota cluster 1 sebanyak 4090 data, anggota cluster 2 sebanyak 1639 data, anggota cluster 3 sebanyak 849 data dan anggota cluster 4 sebanyak 718 data. Terlihat pula bahwa parameter yang paling berpengaruh terhadap proses *clustering* adalah parameter nitrogen oksida (NOx) yang terlihat dari pusat cluster dari parameter tersebut memiliki variansi tinggi. Sementara itu, parameter yang tidak berpengaruh pada proses cluster adalah parameter benzene (C6H6), relative humidity (RH), absolute humidity (AH) dan temperature (T). Hal ini terlihat dari nilai pusat cluster pada parameter tersebut memiliki variansi rendah. Gambar 5 merupakan grafik perbandingan pusat cluster untuk setiap parameter.



Gambar 5 : *Perbandingan pusat cluster*

2. Proses *clustering* menggunakan fuzzy c-means

Proses *clustering* menggunakan fuzzy c-means dilakukan dengan menggunakan *software* Rapid Miner 7.1. Proses ini dilakukan dengan menggunakan operator dengan konfigurasi yang terlihat pada gambar 6.



Gambar 6 : *Operator Rapid Miner fuzzy c-means*

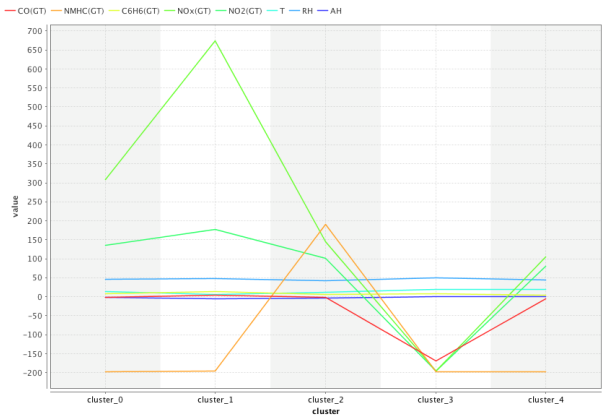
Opearator Read Excel digunakan untuk membaca *spreadsheet* yang berisi data yang sudah melewati tahap pembersihan data. Selanjutnya dilakukan clustering menggunakan *fuzzy* c-*means* pada opeator Fuzzy C-Means. Setelah itu, hasil *clustering* disimpan pada *spreadsheet* dengan penggunakan operator Write Excel.

Proses *clustering* tersebut menggunakan parameter jumlah cluster=5 dan iteration=100 sehingga menghasilkan nilai *centroid* setiap parameter untuk setiap cluster yang tertuang pada tabel 2.

Tabel 2. *Pusat cluster pada fuzzy c-means*

Param	**C0**	**C1**	**C2**	**C3**	**C4**
CO	-1.918	3.215	-1.186	-168.74	-6.3983
NMHC	-196.94	-196.47	191.05	-197.16	-197.02
C6H6	8.3665	13.605	5.1947	7.3517	4.6702
NOx	308.15	673.48	144.24	-195.39	105.49
NO2	135.32	176.78	100.50	-196.24	80.705
T	12.833	6.313	10.732	18.273	18.872
RH	45.759	48.533	41.630	48.978	43.443
AH	-2.142	-5.920	-4.190	-0.598	-0.656

Dengan sebaran anggota cluster 0 sebanyak 2272 data, anggota cluster 1 sebanyak 973 data, anggota cluster 2 sebanyak 873 data, anggota cluster 3 sebanyak 1639 data dan anggota cluster 4 sebanyak 3600 data. Terlihat pada hasil proses *clustering* dengan *fuzzy* c-*means* bahwa parameter yang berpengaruh pada proses *clustering* ini adalah parameter nitrogen oksida (NOx), Non Metanic Hydro Carbon (NMHC) dan natrium dioksida (NO2). Sementara itu, parameter yang tidak berpengaruh pada proses cluster adalah absolute humidity (AH), relative humidity (RH) dan benzene (C6H6). Hal ini terlihat dari nilai pusat cluster pada parameter tersebut memiliki variansi rendah. Gambar 7 merupakan grafik perbandingan pusat cluster untuk setiap parameter.



Gambar 7 : *perbandingan pusat cluster*

3. Pengujian hasil *clustering* dengan standar deviasi

Setelah proses *clustering* dengan menggunakan k-*means* dan *fuzzy* c-*means* dilakukan, selanjutnya dilakukan perhitungan rata-rata dari standar deviasi pada setiap parameter anggota masing-masing cluster. Hasil perhitungan stardar deviasi tertuang pada tabel 3.

Tabel 3. *Rata-rata standar deviasi setiap cluster*

Alg	C0	C1	C2	C3	C4	AVG
KMean	41.276	41.907	34.044	62.263	63.413	48.581
FCM	40.077	62.265	63.004	34.044	39.125	47.036

Berdasarkan hasil pengujian rata-rata standar deviasi yang tertuang pada tabel 3, terlihat bahwa rata-rata standar deviasi pada k-*means clustering* sebesar 48.581 sementara rata-rata standar deviasi pada *fuzzy* c-*means* sebesar 47.036.

5. Kesimpulan

Berdasarkan pembahasan diatas, dapat disimpulkan beberapa simpulan sebagai berikut :

1. Telah dilakukan proses *clustering* data polusi udara harian dengan k-*means* dan *fuzzy* c-*means* yang menunjukan bahwa rata-rata standar deviasi pada hasil *clustering fuzzy* c-*means* lebih kecil dari pada rata-rata standar deviasi pada hasil *clustering* k-*means*.
2. Hasil dari pengujian dua teknik *clustering* tersebut menunjukan bahwa parameter natrium dioksida (NO2), non metal hydro carbon (NMHC) dan natrium oksida (NOx) memiliki pengaruh yang signifikan terhadap proses *clustering*.

Daftar Pustaka

[1] Vorlaufer. 2011. K. Sudostasien (Southeast Asia). 2nd Ed p. 86. Darmstadt.
[2] Wulansari, K, "Evolusi Konsep Ruang Hijau Publik di Kota Semarang pada Awal Abad ke 20 Hingga Sekarang (Ruang Hijau Publik di Kawasan Candi Baru)" dalam Jurnal Pembangunan Wilayah dan Kota Universitas Diponegoro Vol 11 No. 1 Maret 2015.
[3] S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario", Sensors and Actuators B: Chemical, Volume 129, Issue 2, 22 February 2008, Pages 750-757, ISSN 0925-4005.
[4] Yanti, N. Ulfah, M. "Aplikasi Jaringan Syaraf Tiruan Untuk Clustering Polutan Kimia Penyebab Pencemaran Udara" dalam Jurnal Teknologi Terpadu Politeknik Negeri Balikpapan Vol 3 No 2 Oktober 2015.
[5] Rachmatin, D. "Aplikasi Metode Agglomerative dalam Analisis Cluster Pada Data Tingkat Polusi Udara" dalam Jurnal Ilmiah Prodi Matematika STKIP Siliwangi Bandung Vol 3 No 2 September 2014
[6] Sitepu, R., Irmeilyana, Gultom, B., "Analisis Cluster Terhadap Tingkat Pencemaran Udara Pada Sektor Industri di Sumatera Selatan" dalam Jurnal Penelitian Sains Universitas Sriwijaya Vol 14 No. 3 Juli 2011.
[7] Han, J., & Kamber, M., 2006, *Data Mining: Concepts and Techniques 2e*, Morgan Kaufmann Publishers, San Francisco.
[8] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. Elsevier, 2011.
[9] Tan, P. N., Steinbach, M., & Kumar, V., 2006, *Introduction to Data Mining*, Pearson Education, Boston.

ISSN : 2302-3805

Biodata Penulis

Sandi Fajar Rodiyansyah, memperoleh gelar Sarjana Pendidikan (S.Pd.), Pada Program Studi Pendidikan Ilmu Komputer Universitas Pendidikan Indonesia, lulus tahun 2009. Memperoleh gelar *Master of Computer Science* (M.Cs.) Program Pasca Sarjana Magister Ilmu Komputer Universitas Gajah Mada Yogyakarta, lulus tahun 2012. Saat ini menjadi Dosen di Universitas Majalengka – Jawa Barat.

ISSN : 2302-3805