STUDI KOMPARASI METODE KLASTERISASI DATA K-MEANS DAN K-HARMONIC MEANS

I Made Widiartha
Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
email : imadewidiartha@cs.unud.ac.id

Abstrak

Salah satu metode *partitional clustering* yang sangat popular adalah K-Means Clustering (KM). Metode ini banyak digunakan karena implementasinya yang sederhana, dapat menangani data dalam jumlah besar dan proses yang relatif singkat. Meskipun demikian jika diperhatikan dari tahapan KM untuk mendapatkan klaster akhir masih terdapat kelemahan. Titik awal pusat klaster pada KM yang ditentukan secara random dan proses pembaharuan titik pusatnya sangat memungkinkan hasil klaster konvergen pada lokal optimal.

K-Harmonic Means Clustering (KHM) merupakan algoritma yang diciptakan untuk menyempurnakan KM. Dalam KHM titik pusat diperbaharui dengan memanfaatkan rata-rata harmonik dari seluruh titik data ke seluruh pusat klaster yang ada. Rata-rata harmonik dalam metode KHM digunakan untuk mengurangi permasalahan yang ada pada KM. Pada penelitian ini dilakukan studi komparasi terhadap dua metode klasterisasi yaitu KM dan KHM. Penelitian ini ditujukan untuk melihat bagaimana performa metode KHM dalam menyempurnakan metode KM. Studi komparasi ini menggunakan lima buah data set.

Keywords: K-Means Clustering, K-Harmonic Means Clustering, Klasterisasi Data.

Abstract

One of the popular partitional clustering methods is K-Means Clustering (KM). This method is widely used because of its simple implementation, it can handle large amounts of data and processes are relatively short. However, the stage of KM to get the final cluster has some drawbacks. The starting point of the central cluster on KM is determined randomly and its center renewal process allows the cluster converge to a local optimum.

K-Harmonic Means Clustering (KHM) is an algorithm that was created to improve KM. In KHM the center is updated by utilizing the harmonic mean of all data points to all existing cluster centers. Harmonic average of KHM method is used to reduce existing problems in KM. This research conducted a comparative study of two methods, KM and KHM. This study aimed to see the performance of KHM methods in perfecting KM method. This comparative study uses five data sets.

Keywords: K-Means Clustering, K-Harmonic Means Clustering, Data Clustering.

1. Pendahuluan

Pengelompokan data ke dalam beberapa klaster sehingga data dalam satu klaster memiliki tingkat kemiripan yang maksimum dan data antar klaster memiliki kemiripan yang minimum disebut klasterisasi data (*clustering*) [5]. Salah satu metode *partitional clustering* yang sangat popular adalah K-Means Clustering (KM). Dalam proses klasterisasi, metode ini menggunakan teknik partisi yang secara iteratif untuk meminimalkan jarak antara tiap data dengan pusat klasternya. Metode KM dimulai dengan pembentukan titik pusat klaster secara random sehingga didapat prototipe awal klaster yang kemudian secara iteratif pusat klaster ini diperbaiki hingga konvergen (tidak terjadi perubahan yang signifikan pada prototipe klaster). Perubahan ini diukur menggunakan fungsi tujuan yang umumnya didefinisikan sebagai jumlah kuadrat jarak tiap item data dengan pusat klasternya. Metode ini banyak digunakan

karena implementasinya yang sederhana, dapat menangani data dalam jumlah besar dan proses yang relatif singkat. Meskipun demikian jika diperhatikan dari tahapan KM untuk mendapatkan klaster akhir terlihat kelemahan dimana keakuratan hasil klaster sangat tergantung dari penentuan titik awal pusat klaster sehingga permasalahan sensitifitas terhadap penentuan titik awal menjadi kelemahan metode ini. Disamping itu titik awal pusat klaster yang ditentukan secara random sangat memungkinkan hasil klaster konvergen pada lokal optimal [9].

Untuk mengatasi masalah yang terjadi pada inisialisasi pusat klaster, Zhang, Hsu, dan Dayal [11] mengusulkan sebuah metode baru yang diberi nama K-Harmonic Means (KHM) yang kemudian dimodifikasi oleh Hammerly dan Elkan [3]. Tujuan dari metode ini adalah meminimalisasi rata-rata harmonik dari seluruh titik data ke seluruh pusat klaster yang ada. Rata-rata harmonik yang digunakan dalam metode KHM telah terbukti mampu mengurangi permasalahan inisialisasi

2. K Means Clustering

Metode K-Means Clustering (KM) merupakan metode klasterisasi secara partisi (*partitional clustering*). Hampir semua metode klasterisasi secara partisi didasarkan pada tujuan untuk mengoptimalkan nilai fungsi *f(x)* sebagai *clustering criterion* dimana hal ini dapat dikatakan sebagai penerjemahan gagasan intuisi manusia terhadap suatu klaster kedalam suatu rumus matematis (Pen, 1999). KM tidak menjamin hasil klasterisasi yang unik karena metode ini dapat menghasilkan hasil klaster yang berbeda tergantung dari posisi inisialisasi klaster awal (Khan, 2004).

Berikut ini akan diberikan gambaran KM. Misal $X = \{x_i | i = 1, ..., n\}$ merupakan suatau himpunan n titik berdimensi d yang akan diklasterkan kedalam K klaster $C = \{c_k | k = 1, ..., K\}$. Metode KM menemukan suatu partisi/klaster sedemikian hingga nilai *squared error* antara titik tengah (mean) dari suatu klaster ke smua titik data klaster tersebut merupakan nilai minimum (Jain, 2010). Misalkan μ_k adalah rata-rata dari klaster c_k yang didapat dari persamaan 2.6.

$$\mu_k = \frac{1}{n_k} \sum_{x_i \in c_k} x_i$$

Dimana n_k merupakan jumlah elemen pada c_k.*Squared error* antara μ_k dan seluruh data pada klaster c_k didasarkan pada jarak *Euclidean* antara titik yang ada dengan pusat klasternya, *squared error* tersebut didefinisikan sebagai berikut:

$$J(c_k) = \sum_{x_i \in c_k} \| x_i - \mu_k \|^2$$

Fungsi tujuan (*objective function*) dari klasterisasi dengan KM adalah meminimukan total *squared error* dari seluruh klaster. Fungsi tujuan ini juga disebut sebagai *clustering criterion* (Pen, 1999) dan juga sebagai *cost function*(Khan, 2004) dalam penemuan solusi optimal. Adapun formula dari tujuan ini adalah :

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \| x_i - \mu_k \|^2$$

Solusi pada metode KM adalah terbentuknya klaster-klaster dengan nilai $J(C)$ yang minimum. Berikut adalah metode K-Means Clustering :

1. Inisialisasi K titik pusat klaster awal secara acak
2. Klasterkan setiap obyek yang ada sesuai jarak terdekat ke pusat klaster yang ada
3. Perbaiki nilai semua pusat klaster
4. Ulangi langkah 2 dan 3 sampai nilai semua pusat klaster tidak ada perubahan.

3. K-Harmonic Means Clustering

K-Harmonic Means Clustering (KHM) merupakan metode yang diperkenalkan oleh Zhang, Hsu, dan Dayal yang dibuat untuk mengatasi permasalahan yang ada pada K-Means Clustering [10]. KHM merupakan salah satu contoh *center-based cluster* dan merupakan sebuah metode dimana klaster-klaster dibentuk dengan peyempurnaan secara iteratif berdasarkan letak titik pusat dari masing-masing klaster. Pada KHM, nilai fungsi tujuan dihasilkan dengan mencari total rata-rata harmonik dari seluruh titik data terhadap jarak antara masing-masing titik data ke seluruh titik pusat klaster yang ada [10].

Hal ini berbeda dengan KM dimana fungsi tujuan diperoleh dari total jarak seluruh data ke titik pusat klasternya. Rata-rata harmonik ini didefinisikan seperti persamaan 2.1.

$$HA(\{a_i \mid i = 1,..., K\}) = \frac{K}{\sum_{i=1}^{K} \frac{1}{a_i}} \quad (2.1)$$

Dalam fungsi harmonik, jika terdapat satu anggota dalam a_1, ..., a_k bernilai kecil maka nilai rata rata harmonik pun bernilai kecil, tetapi jika tidak ada anggota bernilai kecil maka nilainya pun besar [2].

Rata-rata harmonik sangat sensitif dengan keadaan dimana terdapat dua atau lebih titik pusat yang saling berdekatan. Metode ini secara natural menempatkan satu atau lebih titik pusat ke area titik data yang jauh dari titik-titik pusat yang ada sebelumnya. Hal ini akan membuat fungsi tujuan akan semakin kecil. Adapun langkah-langkah Metode KHM adalah sebagai beriku[8].

1. Inisialisasi posisi titik pusat klaster awal secara random
2. Hitung nilai fungsi tujuan dengan persamaan 2.2, dimana p adalah input parameter. Nilai p biasanya ≥ 2.

$$KHM(X,C) = \sum_{i=1}^{N} \frac{K}{\sum_{l=1}^{K} \frac{1}{\| x_i - c_l \|^p}} \quad (2.2)$$

3. Untuk setiap data x_i, hitung nilai keanggotaan $m(c_l|x_i)$ untuk setiap titik pusat klaster c_l berdasarkan persamaan 2.3.

$$m(c_l \mid x_i) = \frac{\| x_i - c_l \|^{-p-2}}{\sum_{l=1}^{k} \| x_i - c_l \|^{-p-2}} \quad (2.3)$$

4. Untuk setiap data x_i, hitung nilai bobot $w(x_i)$ berdasarkan persamaan 2.4

$$w(x_i) = \frac{\sum_{l=1}^{K} \| x_i - c_l \|^{-p-2}}{\left(\sum_{l=1}^{K} \| x_i - c_l \|^{-p} \right)^2} \quad (2.4)$$

5. Untuk setiap titik pusat c_l, ulang kembali perhitungan untuk posisi titik pusat klaster dari semua data berdasarkan nilai keanggotaan dan bobot yang dimiliki tiap data.

$$c_l = \frac{\sum_{i=1}^{N} m(c_l \mid x_i).w(x_i).x_i}{\sum_{i=1}^{N} m(c_l \mid x_i).w(x_i)} \quad (2.5)$$

6. Ulangi langkah 2 sampai 5 sampai mendapatkan nilai fungsi tujuan yang tidak terdapat perubahan yang signifikan.
7. Tetapkan keanggotaan data x_i pada suatu klaster dengan titik pusat klaster c_l sesuai dengan nilai keanggotaan x_i terhadap c_l.

x_i merupakan anggota dari klaster dengan titik pusat klaster c_l apabila nilai keanggotaan $m(c_l|x_i)$ adalah yang terbesar dibandingkan dengan nilai keanggotaannya ke titik pusat klaster lain..

4 Data Penelitian

Dataset yang digunakan dalam penelitian ini terdiri dari dataset Iris, Wisconsin Breast Cancer (Cancer), Contraceptive Method Choice (CMC), Glass, dan Wine. Data dalam penelitian diambil dari *UCI Machine Learning Repository* (ftp://ftp.ics.uci.edu./pub/machine-learning-databases/). Informasi jumlah fitur, kelas, dan data dapat dilihat pada Tabel 1. Dalam penelitian ini, 80% data akan digunakan sebagai data training dan sisanya digunakan sebagai data testing. Data training ini digunakan untuk melihat performa dari ketiga

Tabel 1. Pembagian Data Set

Dataset	Fitur	Kelas	Jumlah Data	
			Training	Testing
Iris	4	3	120	30
Cancer	9	2	547	136
CMC	9	3	1179	294
Glass	9	6	172	42
Wine	13	3	143	35

metode dalam melakukan klasterisasi data. Penilaian performa ini dilihat dari tiga sudut pandang yaitu nilai fungsi tujuan KHM(X,C), F-Measure, dan waktu eksekusi (*running time*). Data testing hanya digunakan untuk melihat korelasi secara eksternal (kelas label) yaitu bagaimana hasil klasifikasi data testing dengan memanfaatkan hasil titik pusat klaster dengan menggunakan data training.

5. Hasil

Dalam melakukan uji coba pada penelitian ini, nilai parameter yang digunakan untuk metode ABC mengacu pada nilai parameter yang digunakan oleh Zhang. Parameter tersebut antara lain Limit bernilai 10 dan jumlah MCN yang bernilai 2000 [11]. Untuk metode ABC-KHM penentuan, Limit, Max1, Max2 dan Max3 ditentukan dengan melakukan uji coba nilai-nilai parameter ini. Dari hasil uji coba yang telah dilakukan, didapatkan bahwa hasil terbaik diperoleh dengan menggunakan Max1=20, Limit=3, Max2=10, dan MaxABCKHM=20.

Untuk mengetahui performa masing-masing metode maka pada penelitian ini digunakan tiga tolak ukur yaitu nilai fungsi tujuan KHM(X,C), F-measure, dan running time. Uji coba pada penelitian ini dilakukan melalui beberapa sekenario untuk menguji performa dari metode-metode yang ada. Sekenario ini dibuat dengan menggunakan fungsi tujuan yang berbeda-beda. Perbedaan fungsi tujuan ini terletak pada parameter *p*. Pada penelitian ini, terdapat dua buah sekenario nilai *p* yaitu $p = 2$, dan $p = 3$.

Dari sisi penilaian eksternal (kelas label) pada penelitian ini digunakan penilaian F-measure. Nilai F-measure didapat dari persamaan 2.9 [1].

$$F(i,j) = \frac{(b^2+1).(p(i,j).r(i,j))}{b^2.p(i,j)+r(i,j)} \qquad (2.9)$$

$p(i,j) = n_{ij}/n_j$ dan $r(i,j) = n_{ij}/n_i$ dimana n_i adalah jumlah data dari kelas *i* yang diharapkan sebagai hasil query, n_j adalah jumlah data dari klaster *j* yang dihasilkan oleh query, dan n_{ij} adalah jumlah elemen dari kelas *i* yang masuk di klaster *j*. untuk mendapatkan pembobotan yang seimbang antara *precision* dan *recall* maka nilai $b = 1$ digunakan dalam menghitung nilai F-measure [4].

Untuk mendapatkan kesimpulan akhir hasil klasterisasi menggunakan metode-metode yang ada, maka uji coba klasterisasi dilakukan sebanyak 10 kali untuk tiap-tiap sekenario yang dibuat. Kesimpulan kinerja dari metode akan didapatkan melalui nilai rata-rata (mean) dan standar deviasi dari 10 percobaan tersebut.

Seperti yang dibahas pada bagian sebelumnya bahwa data testing hanya digunakan untuk melihat secara eksternal (kelas label) bagaimana hasil klasifikasi data testing dari masing-masing metode klasterisasi. Teknik pengklasifikasian data testing adalah dengan membandingkan jarak antara data tersebut dengan pusat-pusat klaster yang ada. Data testing yang memiliki jarak terdekat dengan suatu titik pusat maka data tersebut akan diklasifikasikan kedalam kelas terdekat dengannya. Hasil uji coba untuk performa setiap metode terhadap tiga tolak ukur dapat dilihat pada Tabel 2 dan 3 .

Tabel 2. Rata-rata dan Standar Deviasi Hasil Uji Coba dengan $p = 2$

Dataset	Pengukuran	Fungsi Tujuan		F-measure		Waktu	
		KHM	KM	KHM	KM	KHM	KM
Iris	Mean	**152,8750**	153,4879	**0,8977**	0,8821	0,08	**0,018**
	Std. Dev.	0,0014	0,0862	0	0,0061	0,0262	0,0054
Cancer	Mean	**24.580,5932**	24.654,5722	0,9503	**0,9587**	0,21	**0,04**
	Std. Dev.	0,0002	4,9	0	0,0010	0,0191	0,0092
Cmc	Mean	**980,3589**	1.059,8764	**0,3925**	0,3554	1,49	**0,15**
	Std. Dev.	2,5412	54,982	0,0158	0,0718	0,6226	0,0405
Glass	Mean	**34,5098**	42,6918	**0,4227**	0,3607	0,40	**0,12**
	Std. Dev.	0,0366	2,7577	0,0061	0,0503	0,0684	0,0195
Wine	Mean	**68,6951**	75,0625	**0,9306**	0,7540	0,13	**0,03**
	Std. Dev.	0,0009	7,2678	0,0062	0,2352	0,0167	0,0057

Tabel 3. Rata-rata dan Standar Deviasi Hasil Uji Coba dengan $p = 3$

Dataset	Pengukuran	Fungsi Tujuan		F-measure		Waktu	
		KHM	KM	KHM	KM	KHM	KM
Iris	Mean	**154,3991**	189,1235	**0,8977**	0,8348	0,08	**0,023**
	Std. Dev.	0,0016	98,6800	0	0,1505	0,0172	0,0055
Cancer	Mean	**208.486,2686**	216.033,1343	0,9387	**0,9583**	0,21	**0,05**
	Std. Dev.	0,0002	118,2	0	0,0009	0,0223	0,0114
Cmc	Mean	**947,3022**	995,1305	**0,3966**	0,3843	1,05	**0,20**
	Std. Dev.	0,0023	46,58	0,0010	0,0151	0,2370	0,0428
Glass	Mean	**21,0367**	26,9837	**0,3726**	0,2207	0,37	**0,12**
	Std. Dev.	0,4727	3,3952	0,0205	0,0848	0,0707	0,0256
Wine	Mean	**46,0725**	47,4959	**0,9375**	0,8935	0,11	**0,04**
	Std. Dev.	1,2847	0,0190	0,0323	0,1465	0,0190	0,008

6. Kesimpulan

Dari hasil penelitian ini didapatkan hasil bahwa metode KHM telah terbukti berhasil mengoptimalkan posisi titik pusat klaster dengan mengarahkan hasil klaster menuju solusi global optimal. Hal ini dibuktikan dengan hasil penelitian yang menunjukkan nilai fungsi tujuan *objective function* dari metode KHM memiliki nilai yang lebih kecil dari metode KM disemua percobaan. Dari sisi penilaian hasil klaster secara eksternal menggunakan F-measure, metode KHM terlihat mendominasi daripada metode KM.

Dari sisi waktu yang dibutuhkan untuk melakukan proses klasterisasi data, metode KHM membutuhkan waktu lebih lama dibandingkan dengan metode KM. Hal ini disebabkan oleh proses dalam KHM yang lebih kompleks daripada proses dalam KM.

7. Referensi

[1] Dalli, A 2003, Adaptation of the F-Measure to cluster-based Lexicon quality evaluation, In EACL, Budapest.

[2] Gungor, Z. dan Unler, A. 2007, "K-Harmonic Means Data Clustering with Simulated Annealing Heuristic", Applied Mathematics and Computation, Vol. 184, hal. 199–209.

[3] Hammerly, G., dan Elkan, C. 2002, "Alternatives to The K-Means Algorithm that Find Better Clusterings", Proceedings of the 11th international conference on information and knowledge management, hal. 600–607.

[4] Handl, J., Knowles, J., dan Dorigo, M. 2003, "On the performance of ant-based clustering. Design and Application of Hybrid Intelligent Systems, Vol. 104, hal. 204–213.

[5] Tan, P.N., Stainbach, M., dan Kumar, V. 2006, Introduction to Data Mining, 4th edition, Pearson Addison Wesley, New York.

[8] Yang, F., Sun, T., dan Zhang, C. 2009, "An Efficient Hybrid Data Clustering Method Based on K-Harmonic Means and Particle Swarm Optimization", Expert Systems with Applications, Vol. 36, hal. 9847–9852.

[9] Pen, J.M., Lozano, J.A., dan Larranaga, P. 1999, "An Empirical Comparison of Four Initialization Methods for The K-Means Algorithm", Pattern Recognition Letters, Vol. 20, hal. 1027-1040.

[10] Zhang, B., Hsu, M., dan Dayal, U. 1999, K-Harmonic Means – A Data Clustering Algorithm, Technical Report HPL-1999-124, Hewlett-Packard Laboratories.

[11] Zhang C., Ouyang, D., dan Ning, J. 2009, "An Artificial Bee colony Approach for Clustering", Expert Systems with Applications, Vol. 37, hal 4761–4767