# Optimization of Initial Centroids for K-means using Simulated Annealing

Ali Ridho Barakbah, Arna Fariza, Yuliana Setiowati
Politeknik Elektronika Negeri Surabaya ITS
ridho@eepis-its.edu, arna@eepis-its.edu, yuliana@eepis-its.edu

**Abstrak**

*Kinerja dari pengklasteran dengan menggunakan metode K-means mempunyai tingkat ketergantungan yang tinggi terhadap penentuan awal titik pusat klaster yang dibangkitkan secara random sehingga seringkali menyebabkan hasil pengklasteran terjebak pada local minima. Paper ini memperkenalkan suatu penerapan algoritma baru untuk mengoptimasi awal titik pusat untuk K-Means dengan menggunakan Simulated Annealing. Simulated Annealing adalah satu algoritma yang mengambil analogi kristalisasi logam untuk pencarian titik optimal. Penggunaan Simulated Annealing untuk permasalahan optimasi dipandang sebagai paradigma pencarian optimasi baru untuk menghindari local minima dan menemukan global optimum. Hasil eksperimen pada paper ini menunjukkan bahwa pemakaian Simulated Annealing untuk membangkitkan initial centroids pada K-means relatif lebih baik dibandingkan dengan beberapa algoritma pengklasteran lainnya.*

**Keywords:** *K-means, Simulated Annealing, initial centroids,optimasi K-means.*

## 1. Pendahuluan

Pengklasteran adalah suatu proses untuk mengklasifikasikan data ke dalam kelompok yang sama yang memiliki tingkat similiritas yang tinggi [1,2,3]. Analisa klaster membentuk klaster yang tepat ketika anggota-anggota dari klaster memiliki tingkat similiritas yang tinggi antar mereka (internal homogeniety) dan tingkat similiritas yang rendah dengan klaster-klaster lainnya (external homogeneity) [4,5]. Pengklasteran merupakan suatu proses untuk memetakan f:$D \rightarrow C$ dari suatu data D=$\{d_1, d_2,\ldots, d_n\}$ ke suatu klaster $C$=$\{c_1, c_2,\ldots, c_n\}$ sesuai tingkat similiritas pada $d_i$. Aplikasi pengklasteran dipakai di berbagai bidang, seperti data mining, pengenalan pola, klasifikasi citra, ilmu-ilmu yang terkait dengan biologi, pemasaran, perencanaan kota, document retrival, dan lain-lain.

Metode pengklasteran yang paling terkenal adalah K-Means yang dikembangkan oleh Mac Queen pada tahun 1967. Metode K-means sangat terkenal karena kemampuannya yang dapat dipakai untuk mengkluster data yang besar, dan juga kemampuannya menangani data outlier.Kesederhanaan K-means membuat algoritma ini banyak digunakan di berbagai bidang. K-means merupakan metode pengklasteran secara partitioning yang memisahkan data ke dalam k kelompok yang berbeda. Dengan melalui partitioning secara iteratif, K-means meminimalkan rata-rata jarak setiap data ke klusternya.

Akan tetapi algoritma K-means sangat sensitif dalam penentuan awal titik pusat klaster. K-means membangkitkan awal titik pusat klaster secara random. Pada saat pembangkitan awal titik pusat yang random tersebut mendekati solusi akhir pusat kluster, K-means mempunyai posibilitas yang tinggi untuk menemukan titik pusat klaster yang tepat. Sebaliknya, jika awal titik pusat tersebut jauh dari solusi akhir pusat kluster, maka besar kemungkinan ini akan menyebabkan hasil pengklasteran yang tidak tepat [6]. Karena pembangkitan awal titik pusat secara random itu, K-means tidak menjamin hasil pengklasteran yang unik [7]. Inilah yang menyebabkan metode K-means sulit untuk mencapai global optimum, akan tetapi hanya local minima [8].

Beberapa metode telah diperkenalkan untuk mengoptimasi awal titik pusat untuk K-means. Duda dan Hart (1973) telah mendiskusikan sebuah metode rekursif untuk menginisialisasi nilai rata-rata yang didapatkan dari keseluruhan data dan secara random dibangkitkan K kali [7]. Bradley dan Fayyad (1998) mengusulkan suatu algoritma yang dapat mengoptimasi awal titik pusat dengan menganalisa distribusi data dan probabilitas kepadatan data [9]. Shehroz dan Ahmad (2004) memperkenalkan suatu metode yang disebut Cluster Center Initialization Algorithm (CCIA) untuk menyelesaikan inisilisasi awal titik pusat untuk K-means [7]. CCIA mengkalkulasi nilai rata-rata dan deviasi standar untuk semua atribut data dan kemudian memisahkan data dengan menggunakan kurva normal ke partisi tertentu. CCIA menggunakan K-means dan density-based multi scale data condesation untuk mengamati similaritas pola data sebelum menemukan awal titik pusat untuk K-means.

Simulated Annealing merupakan algoritma optimasi yang mampu menghasilkan nilai optimal global dan menghindari optima lokal [12]. Simulated Annealing digunakan untuk menentukan awal titik pusat klaster

Ali Ridho Barakbah, Arna Fariza, Yuliana Setiowati, "Optimization of Initial Centroids for K-means using Simulated Annealing",
Proc. Industrial Electronics Seminar (IES) 2005, pp.286-289, November 24, 2005, Surabaya, Indonesia.



pada K-means sehingga diperoleh hasil klaster yang optimal.

## 2. Simulated Annealing

Simulated annealing dikembangkan oleh Kirkpatric (1983) yang digunakan untuk optimasi kombinatorial yang merupakan variant dari algoritma Metropolis. Simulated annealing adalah teknik optimasi numerik dengan prinsip *thermo-dynamic*. Annealing adalah proses dimana material solid dilebur dan didinginkan secara perlahan-lahan dengan mengurangi temperatur. Partikel dari material berusaha menyusun dirinya sendiri selama proses pendinginan. Kumpulan status energi dari partikel yang dibentuk disebut "konfigurasi" dari material. Probabilitas partikel pada semua level energi dapat dihitung dengan menggunakan distribusi Boltzmann. Dengan turunnya temperatur, distribusi Boltzmann mempertahankan konfigurasi partikel yang mempunyai energi terendah. Metropolis menemukan bahwa proses *equilibrium* dapat disimulasikan untuk temperatur tetap dengan menggunakan metode Monte Carlo untuk membangkitkan deretan state energi. Sistem mengalami perubahan state dalam membentuk konfigurasi partikel yang baru. Level energi sebelum perubahan state ($E_s$) dan level energi setelah perubahan state ($E_i$) dibandingkan. Jika $E_s$ lebih besar daripada $E_i$ (berarti $\Delta E > 0$), sistem baru diterima sebagai konfigurasi partikel yang baru. Jika $\Delta E > 0$, probabilitas menerima sistem yang mengalamai perubahan state menggunakan kriteria Metropolis yang disebut faktor probabilitas Boltzmann.

$$p = e^{\left(\frac{-\Delta E}{kT}\right)} \quad (1)$$

dimana:
$k$ = konstanta Boltzmann
T = temperatur
$\Delta E$ = perbedaan level energi sebelum dan sesudah perubahan state

Angka random, $P$, dituliskan sebagai distribusi random uniform pada interval [0,1]. Jika $P > p$, maka perubahan state ditolak dan step baru digunakan untuk posisi saat ini. Jika $p > P$, step perubahan state diterima dan konfigurasi baru menggantikan yang lama. Step baru ditempatkan relatif untuk konfigurasi ini. Kriteria memungkinkan *variable* baru diterima sebagai konfigurasi baru meskipun mempunyai nilai fungsi respon yang jelek dibandingkan konfigurasi saat ini. Memindahkan yang sangat jelek masih kemungkinan diterima daripada memindahkan yang tidak jelek. Gambaran ini merupakan algoritma untuk meninggalkan local optimal. Step baru diambil sampai kriteria terminasi dicapai.

Algoritma Simulated Annealing adalah sebagai berikut:

1. Tentukan Temperatur
2. Bangkitkan State awal (S)
3. Hitung Energi State awal (E)
4. Bangkitkan State baru (S')
5. Hitung Energi State baru (E')
6. S' diterima jika memenuhi Probabilitas Boltzman
7. Lakukan penurunan temperatur
8. Jika belum mencapai equilibrium, kembali ke langkah 3
9. Solusi optimal

## 3. Pemodelan Simulated Annealing

Untuk dapat dipakai untuk mengoptimasi initial centroids, maka peru dilakukan pemodelan Simulated Annealing. Dalam paper ini, kami melakukan pemodelan Simulated Annealing yang meliputi representasi state, energi, penurunan temperatur, probabilitas Boltzman, dan penentuan state baru. Implementasi algoritma Simulated Annealing untuk mengoptimasi initial centroids pada K-means dalam paper ini disebut dengan algoritma pengklasteran SA-Kmeans.

### 3.1. *Representasi State*

Pada penelitian ini state $S$ adalah kombinasi centroid (titik pusat) dari $k$ cluster dengan jumlah $p$ atribut dimana State didefinisikan $S = (X_1,\ldots, X_k)$ dimana $X_i = (x_{i1},\ldots, x_{ip})$.

### 3.2. *Representasi Energi*

Representasi energi pada paper ini menggunakan fungsi tujuan $J$ sebagai representasi suatu cluster yang baik. Pada sebagian besar kasus pengklasteran, untuk menemukan suatu cluster yang baik dapat dilakukan dengan meminimasi fungsi tujuan $J$ [13, 14]. Fungsi tujuan $J$ dapat dideskripsikan sebagai berikut. Katakanlah $X = (x_1,\ldots,x_n)$ dari suatu data clustering dimana $x_i = (x_{i1},\ldots,x_{ip})$. Masing-masing dari data tersebut akan dilakukan pengklasteran dengan $k$ cluster sehingga nantinya akan terbentuk cluster $W = (w_1,\ldots,w_k)$ dari data hasil cluster $w_i = (w_{i1},\ldots, w_{ip})$ dimana $i$=1…$k$. Katakanlah $d(x,w)$ adalah menotasikan metrik jarak dari masing-masing data $x$ ke $w$. Fungsi tujuan $J$ dapat didefinisikan sebagai berikut:

$$J = \sum_{i=1}^{n} \min_r d\left(x_i, w_r\right) \quad (2)$$

### 3.3. *Representasi Penurunan Temperatur*

Representasi penurunan temperatur pada penelitian ini adalah menggunakan jadwal penurunan sebagai berikut:

$$T_i = T_0 x \left(\frac{T_n}{T_0}\right)^{\frac{i}{n}} \quad (3)$$

Ali Ridho Barakbah, Arna Fariza, Yuliana Setiowati, "Optimization of Initial Centroids for K-means using Simulated Annealing",
Proc. Industrial Electronics Seminar (IES) 2005, pp.286-289, November 24, 2005, Surabaya, Indonesia.



### 3.4. *Representasi Probabilitas Boltzman*

Representasi faktor probabilitas Boltzman pada penelitian ini memakai persamaan berikut:

$$p = e^{\left(\frac{-\Delta E}{kT}\right)} \quad (4)$$

dengan $k$=1. Pengesetan k=1 pada paper ini dimaksudkan agar nilai energi di suatu state yang baru itu bernilai sama dengan atau lebih kecil dari energi yang dipunyai oleh state sekarang (*current state*).

### 3.5. *PenentuanState Baru*

Pemodelan state baru penelitian ini dilakukan pada salah satu centroid yang dipilih secara acak. Pemilihan pada salah satu centroid ini dimaksudkan agar perubahan centroid tidak dilakukan secara drastis. Kalau perubahan dilakukan pada semua centroid, maka ini akan memungkinkan terjadinya keadaan state yang diam akibat perubahan energi yang tidak menentu.

Setelah salah satu centroid terpilih, maka selanjutnya dilakukan perubahan nilai secara acak per atribut pada centroid tersebut yang disesuaikan dengan batasan nilai minimum dan maksimum yang ada pada masing-masing atribut.

### 3.6. *Algoritma SA-Kmeans*

Algoritma optimasi initial centroids untuk K-means dengan menggunakan Simulated Annealing adalah sebagai berikut:

1. Tentukan State awal ($S$) dan Energi awal ($E$)
2. Masukkan Temperatur awal ($T_0$) dan Temperatur akhir ($T_n$)
3. Tentukan jumlah iterasi (*itr*)
4. Set $i$ = 1 dan $T_i = T_0$.
5. Tentukan State baru ($S'$) dan ($E'$)
6. Hitung Probabilitas Boltzman ($PB$)
7. Bangkitkan bilangan acak $p$ dimana $0<p<1$.
8. Jika E'<E atau p<PB lakukan update state
9. Lakukan penurunan temperatur ($Ti$)
10. Kembali ke langkah 5 jika $i \leq itr$.
11. Lakukan pengklasteran K-means dengan $S$ sebagai initial centroids

## 4. Hasil Uji Coba

Untuk membuktikan bahwa algoritma pengklasteran SA-Kmeans yang diusulkan dalam paper ini dapat bekerja dengan baik, maka kami serangkaian uji coba. Dalam percobaan, kami melibatkan beberapa data set standar untuk kasus-kasus pengklasteran [14] yang meliputi data set Ruspini, Iris, Fossil, Wine, New Thyroid dan Letter Recognition. Selain itu, untuk mengukur tingkat presisi algoritma SA-Kmeans, maka memakai analisa klaster dengan memakai error ratio [7]. Perumusan dari error ratio adalah sebagai berikut:

$$Error = \frac{Numberofmisclassified}{Numberofpatterns} x100\% \quad (5)$$

Sebagai algoritma pengklasteran pembanding dalam percobaan, kami melibatkan algoritma Single linkage, Centroid linkage, Complete linkage, Average linkage, Fuzzy C-Means dengan derajat kefuzzyan=1.25 dan K-means dengan inisialisasi random dengan mengambil rata-rata dari 100 percobaan. Untuk beberapa data set, kami menambahkan Cluster Center Initialization Algorithm (CCIA) [7] sebagai algoritma pembanding. Juga untuk data set Letter Recognition, paper ini membatasi hanya pada pengenalan 595 pola huruf A dan 597 pola huruf D sebagai pembanding terhadap CCIA. Dalam percobaan, kami memang sengaja tidak melakukan normalisasi data dengan tujuan agar dapat fokus untuk menguji kinerja algoritma clustering.

| Algoritma | Error ratio | | | | | |
|---|---|---|---|---|---|---|
| | Ruspini | Iris | Fossil | Wine | New Thyroid | Lett. Rec. |
| Single linkage | 0 | 32 | 13,7931 | 57,3034 | 29,7674 | 49,8322 |
| Centroid linkage | 0 | 9,3333 | 11,4943 | 38,764 | 27,907 | 48,1544 |
| Complete linkage | 4 | 16 | 14,9425 | 32,5843 | 28,3721 | 42,7852 |
| Average linkage | 0 | 9,3333 | 9,1954 | 38,764 | 26,0465 | 6,8792 |
| FCM | 0 | 13,324 | 11,5057 | 30,3371 | 14,4186 | 13,1711 |
| K-means | 13,7787 | 17,7307 | 8,5931 | 32,6197 | 20,9842 | 8,2326 |
| CCIA | 4 | 11,33 | 0 | - | - | 8,55 |
| SA-Kmeans | 0 (100 itr) | 10,6667 (10000 itr) | 25,2874 (10000 itr) | 29,7753 (100 itr) | 15,814 (10000 itr) | 8,2215 (100 itr) |

**Tabel 1.** Perbandingan error ratio pada beberapa algoritma clustering

Tabel 1 menunjukkan error ratio dari beberapa metode pengklasteran untuk semua data set. Terlihat pada tabel tersebut, selain pada data set Fossil, algoritma SA-Kmeans mempunyai kinerja yang relatif lebih baik dibandingkan algoritma-algoritma clustering yang lain untuk data set Ruspini, Iris, Wine, New Thyroid dan Letter Recognition. Bahkan pada data set Ruspini, Wine dan Letter Recognition, jumlah iterasi pada Simulated Annealing hanya 100 kali. Ini menunjukkan bahwa Simulated Annealing berhasil mencapai optimal dengan relatif cepat dan berhasil mengoptimasi initial centroids

untuk K-means sehingga menghasilkan error ratio relatif rendah dibandingkan dengan algoritma-algoritma pembanding.

**5. Kesimpulan**

Paper ini mengangkat masalah optimasi penentuan initial centroids pada K-means dengan menggunakan Simulated Annealing. Algoritma SA-Kmeans yang diajukan pada paper ini menunjukkan kinerja yang relatif lebih baik daripada algoritma-algoritma pengklasteran pembanding. Pada serangkaian uji coba, SA-Kmeans menghasilkan error ratio yang relatif lebih rendah dibandingkan algoritma-algoritma yang lain pada data set Ruspini, Iris, Wine, New Thyroid dan Letter Recognition. Bahkan pada beberapa data set, Simulated Annealing berhasil mencapai optimal dengan cepat berhasil mengoptimasi initial centroids untuk K-means.

## References


[1] G. Karypis, E.H. Han, V. Kumar, Chameleon: a hierarchical clustering algorithm using dynamic modeling, IEEE Computer: Special Issue on Data Analysis and Mining 32(8):68W5, 1999.

[2] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Clustering algorithms and validity measures, proceedings of the 13th International Conference on Scientific and Statistical Database Management, July 18–20. IEEE Computer Society, George Mason University, Fairfax, Virginia, USA, 2001.

[3] S. Bandyopadhyay, An automatic shape independent clustering technique, Machine Intelligence Unit, Journal of Pattern Recognition Society, volume 37, number 1, January 2004.

[4] G.A. Growe, Comparing algorithms and clustering data: components of the data mining process, thesis, department of Computer Science and Information Systems, Grand Valley State University, 1999.

[5] V. Estivill-Castro, Why so many clustering algorithms-a position paper, ACM SIGKDD Explorations Newsletter, Volume 4, Issue 1, pp. 65-75, 2002.

[6] Y.M. Cheung, k*-Means: A new generalized k-means clustering algorithm, Pattern Recognition Lett. 24 (2003) 2883-2893.

[7] S.S. Khan, A. Ahmad, Cluster center initialization algorithm for K-means clustering, Pattern Recognition Lett. (Accepted), 2004.

[8] B. Kövesi, J.M. Boucher, S. Saoudi, Stochastic K-means algorithm for vector quantization, Pattern Recognition Lett. 22 (2001) 603-610.

[9] P.S. Bradley, U.M. Fayyad, Refining initial points for K-means clustering, Proc. 15th Internat. Conf. on Machine Learning (ICML'98), 1998.

[10] B. Kövesi, J.M. Boucher, S. Saoudi, Stochastic K-means algorithm for vector quantization, Pattern Recognition Lett. 22 (2001) 603-610.

[11] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, "Equation of State Calculations by Fast Computing Machines", J. Chem. Phys., vol. 21, no. 6, pp. 1087-1092, 1983.

[12] S.Kirkpatrick, C. D. Gelatt Jr, M. P. Vecchi, "Optimization by Simulated Annealing", Science, 220, 4598, pp. 671-680, 1983.

[13] A. Likas, A reinforcement learning approach to on-line clustering, Department of Computer Science, University of Ioannina, Greece.

[14] UCI Repository (http://www.sgi.com/tech/mlc/db/).

[15] W. Xiao-Ying, M. Jonathan, Garibaldi, Simulated Annealing Fuzzy Clustering in Cancer Diagnosis, International Journal of Computing and Informatics, Volume 29 Number 1, May 2005.