

# Algoritma Cluster Dinamik untuk Optimasi Cluster pada Algoritma K-Means dalam Pemetaan Nasabah Potensial

Widiarina

Magister Ilmu Komputer, STMIK Nusa Mandiri  
Email:widiarina11@gmail.com

Romi Satria Wahono

Fakultas Ilmu Komputer, Universitas Dian Nuswantoro  
Email: romi@brainmatics.com

**Abstract:** Pelanggan merupakan salah satu sumber keuntungan perusahaan. Pemahaman yang baik tentang pelanggan sangat penting dilakukan untuk mengetahui nilai potensial pelanggan. Saat ini pelaksanaan CRM (*Customer Relationship Management*) dapat membantu dalam pemahaman nilai pelanggan. Segmentasi pelanggan adalah salah satu metode yang digunakan untuk pemetaan pelanggan. Nilai potensial pelanggan dapat diukur menggunakan metode RFM (*Recency, Frequency, Monetary*). Algoritma *K-means* salah satu metode yang banyak digunakan untuk segmentasi pelanggan. *K-means* banyak dipakai karena algoritma nya mudah dan sederhana, tetapi algoritma ini memiliki kekurangan yaitu sensitifitas pada partisi awal jumlah cluster(k). Untuk menyelesaikan masalah sensitifitas partisi awal jumlah cluster pada algoritma *K-means*, maka diusulkan algoritma cluster dinamik untuk menetapkan jumlah cluster(k). Hasil percobaan menunjukkan bahwa algoritma cluster dinamik pada *K-means*, dapat menghasilkan kualitas cluster yang lebih optimal.

**Keywords:** segmentasi pelanggan, RFM, K-Means, quality Cluster

## 1 PENDAHULUAN

Pelanggan menduduki posisi penting dalam pengembangan strategi bisnis, pelanggan juga merupakan salah satu sumber keuntungan dalam perusahaan. Untuk itu diperlukan suatu pemahaman yang baik tentang pelanggan. Pemahaman yang baik terhadap pelanggan dapat digunakan perusahaan untuk berinvestasi pelanggan yang potensial. Masalah yang sering dihadapi adalah kesulitan dalam menganalisa nilai pelanggan (Xing, 2010). Banyak pemasar mengalami kesulitan untuk mengidentifikasi pelanggan atau nasabah yang tepat (Chai & Chan, 2008), hal tersebut dapat mengakibatkan perusahaan kehilangan nasabah potensial dan tentunya akan sangat merugikan perusahaan.

Segmentasi pelanggan adalah metode yang populer yang digunakan untuk memilih pelanggan atau nasabah yang tepat untuk memulai promosi (Chai & Chan, 2008). Dengan segmentasi nasabah berdasarkan prilakunya, kita dapat menargetkan tindakan mereka dengan lebih baik. Seperti peluncuran produk yang disesuaikan, target pemasaran dan untuk memenuhi harapan pelanggan (Balaji & Srivatsa, 2012). Namun untuk menganalisa data pelanggan atau nasabah dalam jumlah besar memerlukan tenaga dan waktu yang banyak.

Beberapa algoritma segmentasi telah digunakan dalam segmentasi pelanggan termasuk metode *Self Organizing Map* (Chan, 2005), dan *K-means* (Prasad & Malik, 2011). Algoritma *K-means* adalah algoritma *clustering* yang paling populer digunakan karena memiliki kelebihan yaitu algoritmanya

sederhana, mudah di implementasikan. Dan merupakan salah satu metode yang cukup efisien dalam hal kompleksitas nya  $O(nkt)$  (Aggarwal & Aggarwal, 2012). Salah satu kekurangan algoritma *K-means* yaitu mempunyai masalah sensitifitas terhadap penentuan partisi awal jumlah *cluster(k)* dan solusi akhir menyatu pada *local minima*. Penentuan partisi jumlah *cluster(k)* sangat penting bagi algoritma *K-means*, tetapi tidak ada ketentuan yang berlaku untuk menentukan berapa jumlah *cluster(k)* yang akan dibentuk (Zhang & Fang, 2013). Pada prakteknya penentuan partisi awal jumlah *cluster* sangat sulit, karna penentuan partisi awal jumlah k yang berbeda dapat menghasilkan kelompok *cluster* yang berbeda pula.

Pada penelitian ini kami mengusulkan algoritma *cluster* dinamik pada algoritma *K-means* dalam menetapkan jumlah *cluster(k)* agar dapat menghasilkan kualitas *cluster* yang optimal sehingga memberikan hasil pemetaan nasabah potensial lebih baik dan tepat.

## 2 PENELITIAN TERKAIT

*K-means* merupakan suatu algoritma pengklasteran yang cukup sederhana yang mempartisi dataset kedalam beberapa kluster *k*. Algoritmanya cukup mudah untuk diimplementasi dan dijalankan, relatif cepat, mudah disesuaikan dan banyak digunakan (Wu & Kumar, 2009). Prinsip utama dari teknik ini adalah menyusun *k* buah partisi/pusat massa (*centroid*)/rata-rata (*mean*) dari sekumpulan data. Algoritma *K-means* dimulai dengan pembentukan partisi kluster di awal kemudian secara iteratif partisi kluster ini diperbaiki hingga tidak terjadi perubahan yang signifikan pada partisi kluster (Witten, Eibe, & Hall, 2011).

Beberapa penelitian yang telah dilakukan mengenai masalah sensitifitas inisialisasi jumlah *cluster(k)*, dan algoritma yang digunakan. Penelitian yang dilakukan oleh (Dealers & Auwatanamongkol, 2007) mengusulkan sebuah algoritma partisi data untuk menghitung awal pusat kluster. Partisi data mencoba membagi ruang data kedalam sel kecil atau kelompok, mana yang jarak intercluster sebesar mungkin dan jarak intracluster sekecil mungkin. Sel dipartisi satu persatu sampai jumlah sel sama dengan jumlah kluster (*k*) yang telah ditetapkan, dan pusat-pusat sel *k* menjadi awal pusat kluster untuk *K-means*. Hasil percobaan menunjukkan bahwa algoritma partisi data bekerja lebih baik dibandingkan dengan inisialisasi pusat kluster secara acak dari sebagian kasus eksperimental dan dapat mengurangi waktu *running* algoritma *K-means* untuk dataset yang besar. (Yi et al., 2010), mengusulkan sebuah algoritma partisi data untuk memperbaiki awal pusat *cluster* yaitu Algoritma awal pusat *cluster* berbasis kepadatan (*density*). Algoritma ini menggunakan fungsi gaussian untuk memenuhi konsistensi global fitur *clustering*.

Algoritma yang diusulkan memilih titik kepadatan terbesar sebagai titik pusat awal pertama dari dataset, kemudian menentukan pusat awal kedua menggunakan metode yang sama dari dataset sehingga menghapus titik pertama dan tetangganya. Proses ini berlanjut sampai set  $M$  awal berisi  $k$  poin. Hasil percobaan menunjukkan bahwa algoritma yang diusulkan sangat meningkatkan kualitas dan stabilitas algoritma  $K$ -means. (Zhang & Fang, 2013) melakukan penelitian dalam perbaikan algoritma  $K$ -means untuk mengoptimalkan inisialisasi pusat *cluster*. Dengan menemukan satu set data yang mencerminkan karakteristik distribusi data sebagai pusat awal *cluster* untuk mendukung pembagian data ke batas yang terbaik. Hasil percobaan didapatkan hasil akurasi algoritma perbaikan  $K$ -means meningkat secara signifikan dibandingkan dengan algoritma  $K$ -means tradisional, dan algoritma yang diusulkan menunjukkan bahwa hasil setiap *cluster* lebih kompak.

Pada penelitian ini kami mengkombinasikan algoritma *cluster* dinamik dengan algoritma  $K$ -Means untuk menghasilkan kualitas *cluster* yang optimal dalam segmentasi nasabah potensial.

### 3 METODE YANG DIUSULKAN

Algoritma *cluster* dinamik pada algoritma  $K$ -means dinamik dapat dilihat pada Gambar 1. Algoritma yang diusulkan mencari jumlah *cluster* yang dijalankan berdasarkan kualitas *cluster* keluaran. Diawal cara kerja sama dengan algoritma  $k$ -means, diakhir akan dilakukan perhitungan intra dan inter cluster, jika jarak intra lebih kecil dan jika jarak intra lebih besar, maka algoritma menghitung *cluster* baru dengan menambahkan *counter*  $k$  dengan satu atau  $k=k+1$  disetiap iterasi sampai memenuhi batas validitas kualitas *cluster* yang berkualitas (M & Hareesha, 2012).

Berikut tahapan Algoritma  $K$ -means:

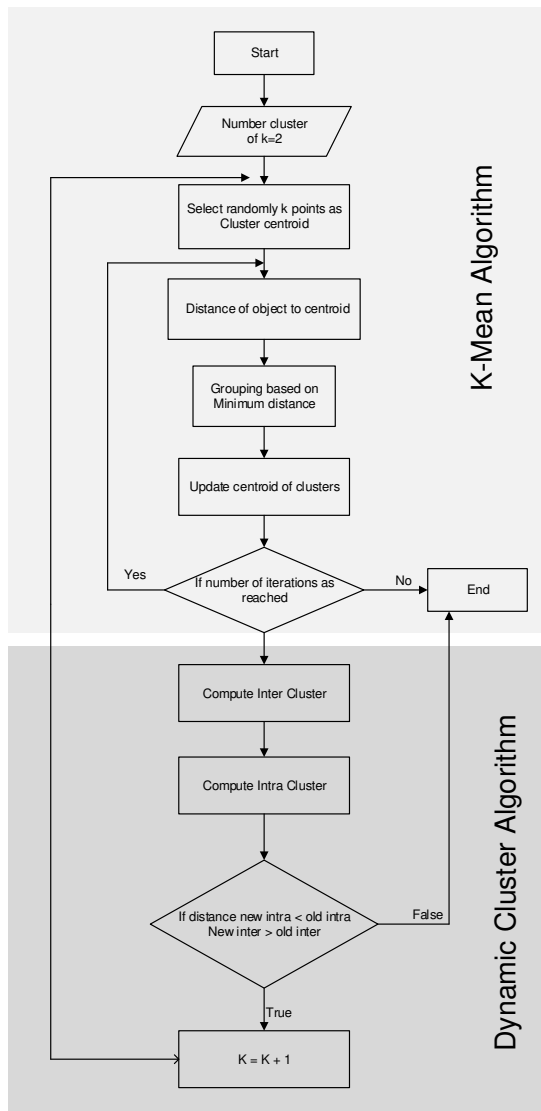
1. Membuat partisi sejumlah  $k$  dari segmentasi yang akan dibentuk.
2. Pilih secara acak  $k$  point untuk dijadikan pusat *cluster*
3. Menghitung jarak data yang lain dengan pusat *cluster*
4. Mengisi setiap obyek dalam dataset kedalam segmen terdekat.
5. Kalkulasi ulang setiap segmentasi yang terbentuk
6. Ulangi langkah hingga data di dalam segmentasi tidak berubah

Istilah inter adalah minimum jarak antar pusat *cluster*, inter digunakan untuk mengukur pemisahan antar *cluster*, yang didefinisikan sebagai:

$$inter = \min \{ \|m_k - m_{kk}\| \} \quad \forall k = 1, 2, \dots, K - 1 \text{ dan } k = k + 1, \dots, K \dots \dots \dots (1)$$

Istilah intra digunakan untuk mengukur kekompakan dari suatu kelompok. Standar deviasi digunakan untuk memeriksa kedekatan titik data setiap *cluster*, dan dihitung sebagai:

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - X_m)^2} \dots \dots \dots (2)$$



Gambar 1. Algoritma *Cluster* Dinamik pada  $K$ -Means

### 4 HASIL PENELITIAN

Dalam melakukan proses perhitungan, pada penelitian ini digunakan komputer dengan spesifikasi komputer Pentium(R) Dual-Core 2.00GHz, RAM 2 GB, Sistem Operasi Windows 7, 32 bit, dan program aplikasi menggunakan *software Matlab*. Dalam penelitian ini, data diperoleh dari data nasabah pengguna jasa penerimaan transaksi kartu pada suatu perusahaan perbankan. Untuk membantu pemilihan atribut, data yang digunakan adalah data nasabah dan data transaksi nasabah selama 6 bulan kebelakang, data awal terdiri dari bulan Juli sampai Desember 2012 dengan jumlah data sebanyak 1002 *record*. Data yang sudah terkumpul akan diolah melalui beberapa tahap pengolahan awal data (*preparation data*). Tahapan pengolahan data yang dilakukan yaitu validasi data, transformasi data, dan seleksi atribut. Contoh data yang sudah di transformasi data dapat dilihat pada Tabel 1. Atribut yang digunakan yaitu :*Last transaction(R)*, *cust age(F)*, dan *rate amount(M)* ketiga atribut tersebut dipilih berdasarkan faktor *Recency Frequency Monetary*.

Table 1. Data Transformasi

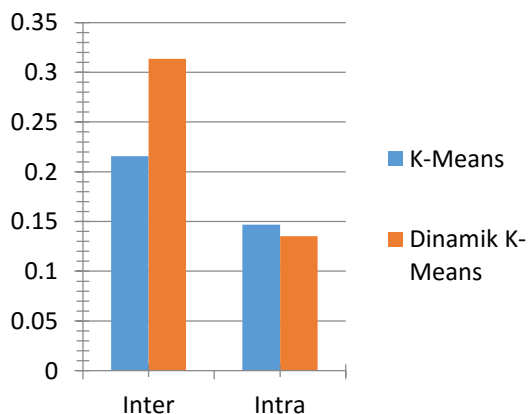
FIELD	Rec	Rec	Rec	Rec	Rec
	1	2	3	4	5
ITEM DEBIT1	4	1	19	99	4
AMOUNT DEBIT1	3	1	16	99	4
ITEM PREPADE1	6	1	7	99	1
AMOUNT PREPADE1	3	1	6	99	1
ITEM CREDIT1	1	1	53	99	23
AMOUNT CREDIT1	1	1	40	99	58
ITEM TOTAL1	4	1	21	99	6
AMOUNT TOTAL1	2	1	18	99	10
LAST TRANSACTION	1	99	80	1	1
CUST AGE	99	87	47	8	1
RATE AMOUNT	2	1	3	99	9

Hasil segmentasi yang terbentuk akan dievaluasi menggunakan *Davies-Bouldin(DB) Index* dan *purity*. *Davies-Bouldin Index* adalah fungsi rasio dari jumlah antara *cluster scatter* sampai dengan *cluster separation* (Maulik & Bandyopadhyay, 2002). *Davies-Bouldin Index* merupakan metode validasi *cluster* dari hasil *clustering*. Pendekatan pengukuran DBI yaitu memaksimalkan jarak *inter cluster* serta meminimalkan jarak *intra cluster*. Nilai *purity* adalah kesesuaian antara *cluster* dengan *cluster* ideal, semakin besar nilai *purity* (semakin mendekati 1), semakin baik kualitas *cluster* (Yi, Qiao, Yang, & Xu, 2010). Semakin kecil nilai DB *Index* menunjukkan skema *cluster* yang paling optimal. Semakin besar nilai *purity* (mendekati 1) semakin baik kualitas *cluster*.

Dari hasil percobaan yang dilakukan, algoritma *K-means* dengan *cluster* dinamik dapat menghasilkan kualitas *cluster* yang lebih baik dibandingkan dengan *K-means* tradisional. Tabel 2 menunjukkan bahwa algoritma *K-means* dengan algoritma *cluster* dinamik menghasilkan jarak *inter* yang lebih besar. Dan jarak *intra* yang lebih kecil dibandingkan dengan jarak *inter* dan *intra* yang dihasilkan oleh algoritma *K-means* tradisional. Dengan peningkatan akurasi yang cukup signifikan, grafik dapat dilihat pada Gambar 2.

Tabel 2. Hasil Percobaan

Algoritma	Cluster(k)	Inter	Intra
K-Means	3	0,2157	0,1469
Dinamik K-Means	4	0,3135	0,1352



Gambar 2. Grafik Hasil Percobaan

Akhirnya, hasil evaluasi *cluster* menunjukkan bahwa algoritma *K-means* dengan algoritma *cluster* dinamik memperoleh nilai DBI lebih kecil dibandingkan dengan algoritma *K-means* tradisional dengan nilai DBI sebesar 0.4313, dan nilai *purity* lebih besar yaitu 0.7647. Hasil evaluasi *cluster* dapat dilihat pada Tabel 3. Nilai DBI yang lebih kecil dan nilai *purity* yang lebih besar, dengan demikian menunjukkan bahwa skema *cluster* lebih optimal.

Tabel 3. Nilai pengujian *Davies-Bouldin Index* dan *purity*

Algoritma	Cluster(k)	DBI	Purity
K-Means	3	0,6810	0,5294
Dinamik K-Means	4	0,4313	0,7647

Segmen yang terbentuk berdasarkan faktor *recency(R)*, *frequency(F)* dan *monetary(M)*. Semakin besar nilai R menunjukkan bahwa nasabah sering melakukan transaksi, semakin besar nilai F menunjukkan bahwa nasabah tersebut setia terhadap terhadap produk yang digunakan, dan semakin besar nilai M, menunjukkan bahwa nilai transaksi yang dibayarkan semakin besar.

1. *Segmen 1*: 331 nasabah, memiliki rata rata nilai yang cukup besar dari ketiga faktor. Maka dapat digolongkan sebagai nasabah yang cukup potensial
2. *Segmen 2*: 44 nasabah, segmen ini tergolong sebagai nasabah yang tidak potensial, karna hanya faktor *recency* saja yang sangat tinggi, sedangkan kedua faktor lainnya sangat kecil.
3. *Segmen 3*: 93 nasabah, dengan nilai dari ketiga faktor yang besar, maka segmen ini tergolong sebagai nasabah yang potensial.
4. *Segmen 4*: 191 nasabah, memiliki nilai yang sangat besar dari ketiga faktor, sehingga segmen ini tergolong sebagai nasabah yang sangat potensial.

### 5 KESIMPULAN

Dari penelitian yang dilakukan, *K-means* dengan algoritma *cluster* dinamik, terbukti dapat meningkatkan akurasi model yang terbentuk. Peningkatan kualitas model dapat dilihat dari peningkatan akurasi yang cukup signifikan. Pengukuran *validity cluster* dengan menggunakan *Davies-Bouldin Index* (DBI) dan *purity*, membuktikan bahwa *K-means* dengan algoritma *cluster* dinamik menghasilkan kualitas *cluster* yang lebih optimal yang ditunjukkan dengan nilai DBI yang lebih kecil dibandingkan dengan *K-means* tradisional, dan *purity* untuk *K-means* dengan algoritma *cluster* dinamik yang lebih besar dibandingkan dengan *K-means* tradisional. Nilai DBI yang lebih kecil mendekati 0 dan *purity* yang lebih besar mendekati 1, menunjukkan skema *cluster* yang optimal.

Meskipun model yang diusulkan sudah memberikan hasil yang lebih baik, namun untuk penelitian selanjutnya dapat menerapkan algoritma Dinamik K-means untuk dataset yang lebih beragam dan berbeda, dan pengurangan waktu komputasi untuk dataset yang besar.

## REFERENSI

- Aggarwal, N., & Aggarwal, K. (2012). Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data mining, *International Journal of Scientific & Eninnering Research*, 3(3).
- Balaji, S., & Srivatsa, S. K. (2012). Customer Segmentation for Decision Support using Clustering and Association Rule based approaches, *International Journal of Computer Science & Engineering Technology*, 3(11), 525–529.
- Chai, C., & Chan, H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer, *Expert System with Application*, 34, 2754–2762.
- Chan, C. H. (2005). Online Auction Customer Segmentation Using a Neural Network Model, *International Journal of Applied Science and Engineering* 101–109.
- Deelers, S., & Auwatanamongkol, S. (2007). Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance, *World Academy of Science, Engineering and Technology*, 26(December), 323–328.
- M, A. S. B., & Hareesha, K. S. (2012). Dynamic Clustering of Data with Modified K-Means Algorithm, 27(Icicn), 221–225.
- Maulik, U., & Bandyopadhyay, S. (2002). Performance Evaluation of Some Clustering Algorithms and Validity Indices, *IEEE Transaction On Pattern Analysis And Machine Intelligence*, 24(12), 1650–1654.
- Prasad, P. (2011). Generating Customer Profiles for Retail Stores Using Clustering Techniques, *International Journal on Computer Science and Enginnering*, 3(6), 2506–2510.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning and Tool*. Burlington: Morgan Kaufmann Publisher.
- Wu, Xindong & Kumar, Vipin. (2009). *The Top Ten Algorithms in Data Mining*. London: CRC Press.
- Xing, B. I. (2010). The Evaluation of Customer Potential Value Based on Prediction and Cluster Analysis, *International Conference on Management Science & Engineering 17th*, Melbourne, Australia 613–618.
- Yi, B., Qiao, H., Yang, F., & Xu, C. (2010). An Improved Initialization Center Algorithm for K-Means Clustering. 2010 *International Conference on Computational Intelligence and Software Engineering*, *IEEE* (1), 1–4.
- Zhang, C., & Fang, Z. (2013). An Improved K-means Clustering Algorithm Traditional K-mean Algorithm, *Journal of Information & Computational Science*, 1, 193–199

## BIOGRAFI PENULIS



**Widiarina.** Memperoleh gelar M.Kom dari Sekolah Tinggi Manajemen Ilmu Komputer Nusa Mandiri, Jakarta. Staf pengajar di salah satu Perguruan Tinggi Swasta. Minat penelitian saat ini pada bidang data mining.



**Romi Satria Wahono.** Memperoleh gelar B.Eng dan M.Eng masing-masing di Ilmu Komputer dari Saitama University, Jepang, dan Ph.D dalam Rekayasa Perangkat Lunak dari Universiti Teknikal Malaysia Melaka. Pengajar di Pascasarjana Ilmu Komputer, Universitas Dian Nuswantoro, Indonesia. Juga pendiri dan CEO PT Brainmatics, sebuah perusahaan pengembangan perangkat lunak di Indonesia. Minat penelitiannya saat ini meliputi rekayasa perangkat lunak dan machine learning. Anggota profesional dari asosiasi ilmiah ACM, PMI dan IEEE Computer Society.