PRISMA 4 (2021): 640-644
PRISMA, Prosiding Seminar Nasional Matematika
https://journal.unnes.ac.id/sju/index.php/prisma/
ISSN 2613-9189

Algoritme *Quick RObust Clustering using linKs* (QROCK) untuk *Clustering* Data Kategorik

Intan Arum Sari[a,*], Dewi Retno Sari Saputro[b]

[a,b] *Program Studi Matematika FMIPA Universitas Sebelas Maret, Jl. Ir Sutami No.36 A, Surakarta 57126, Indonesia*

** Alamat Surel: intanarum_sari@student.uns.ac.id*

Abstrak

Data *mining* adalah proses ekstraksi dari suatu informasi pengetahuan atau pola yang bersifat penting dan bermanfaat dalam suatu basis data yang memiliki ukuran besar dan merupakan bagian integral dari *Knowledge Discovery in Databases* (KDD). *Clustering* merupakan salah satu proses data *mining* yang digunakan untuk melihat pola pendistribusian data yang akan digunakan sesuai dengan karakteristik data. Metode *clustering* hirarki dan non hirarki dikatakan tidak cocok digunakan untuk data kategorik sehingga dikembangkan Metode *RObust Clustering using linKs* (ROCK) dan berikutnya *Quick* ROCK. ROCK, metode yang di dalamnya membangun *link* untuk menggabungkan *cluster-cluster*-nya dan memiliki *robustness* yang baik sehingga dapat mengelompokkan data secara efektif. Sementara QROCK menghitung cluster dengan menentukan komponen yang terhubung dari grafik, diduga hal ini lebih efisien untuk mendapatkan cluster yang memberikan pengurangan waktu komputasi dibandingkan algoritme ROCK. Tujuan penelitian ini untuk mengkaji algoritme QROCK dari aspek akurasi dan efisiensinya. Hasil menunjukkan bahwa *Quick RObust Clustering using linKs* (QROCK) lebih efisien dan lebih akurat karena dapat mendeteksi *outlier* pada data kategorik.

Kata kunci:

clustering, data kategorik, ROCK, QROCK, akurasi, *outlier*

1. Pendahuluan

Knowledge Discovery in Database atau sering disebut dengan KDD merupakan suatu peroses untuk menemukan informasi yang berguna serta pola yang termuat dalam data (Goharian & Grossman, 2003). Informasi yang digunakan termuat dalam basis data yang memiliki ukuran besar, yang sebelumnya tidak diketahui dan bersifat potensial (Han *et al.*, 2011). Salah satu tahapan dari proses KDD adalah data *mining*. Data *mining* adalah proses pengumpulan dan pemakaian data historis untuk menemukan pola keteraturan dalam data dengan jumlah besar (Mujiasih, 2011). Data *mining* dibagi menjadi 6 kelompok yaitu deskripsi, estimasi, prediksi, klasifikasi, asosiasi, dan *clustering* (pengelompokan).

Analisis *clustering* merupakan salah satu teknik analisis multivariat. Menurut He *et al*. (2005), analisis *clustering* adalah proses pengelompokkan beberapa objek yang berdasarkan pada similaritas atau ketidaksamaan 3 atribut dari objek. Pengelompokkan objek dari algoritme *clustering* pada umumnya terfokus pada data numerik atau kontinu, namun beberapa penelitian algoritme *clustering* juga menggunakan data kategorik. Dewangan *et al*. (2010), melakukan penelitian tentang *clustering* untuk data kategorik, yaitu dengan mentransformasikan variabel kategorik ke dalam bentuk numerik, kemudian dari hasil transformasi data dilakukan pengelompokan objek dengan metode fuzzy *clustering*. Kemudian, Ariawan (2019), mendeteksi outlier berbasis kluster pada set data dengan atribut campuran numerik dan kategorikal dengan menggunakan algoritme squeezer. Algoritme squeezer merupakan metode *clustering* yang efektif digunakan dalam data yang berjumlah besar (Johnson & Wichern, 2002). Hanya beberapa algoritme *clustering* yang memiliki ketahanan (*robustness*) yang baik, salah satunya adalah algoritme *RObust Clustering using linKs* (ROCK).

To cite this article:
Sari, I. A., & Saputro, D. R. S. (2021). Algoritme *Quick RObust Clustering using linKs* (QROCK) untuk *Clustering* Data Kategorik. *PRISMA, Prosiding Seminar Nasional Matematika 4*, 640-644

Algoritme ROCK adalah algortime yang digunakan untuk data yang jumlahnya relatif sedikit. Pengembangan penelitian dari algoritme ROCK oleh Guha *et al.* (2000) dengan melakukan *clustering* untuk data kategorik. Dalam penelitian Guha *et al.* (2000) metode ROCK memiliki hasil dengan akurasi yang lebih baik daripada metode hirarki, selain itu, metode ini juga menunjukkan sifat skalabilitas yang baik. Salah satu kelemahan algoritme ROCK adalah tidak bekerja baik untuk data yang mengandung outlier, sehingga *clustering* yang dilakukan tidak dapat akurat. Outlier merupakan data yang memiliki sifat berbeda dan tidak mengikuti tingkah laku umum dari data lainnya, perbedaan yang penting atau sesuatu yang tidak konsisten dalam himpunan data (Kantardzic, 2011). Kemudian Dutta *et al*. (2005) melakukan penelitian tentang algoritme QROCK, yaitu dengan membuat versi cepat dari algoritme ROCK untuk data kategorik. Dalam penelitiannya waktu iterasi yang dibutuhkan algoritme QROCK lebih efisien, selain itu algoritme QROCK juga dikatakan dapat mendeteksi outlier (Guha, 1999). Algoritme QROCK pada dasarnya mempunyai prinsip yang sama dengan algoritme ROCK (Alamsyah, 2006). Menurut Guha (1999), jika dua obyek merupakan *neighbor* maka minimal pasangan obyek itu mempunyai nilai *links* 1 sehingga proses pengelompokan dengan algoritme QROCK dapat dilakukan. Oleh karena itu, pada penelitian ini dilakukan kajian terhadap algoritme QROCK dari aspek akurasi dan efisiensinya.

2. Kajian Teori dan Pembahasan

Data *mining* merupakan proses pembentukan pola dalam suatu data yang memiliki ukuran besar. Selain itu data ming dapat dikatakan juga sebagai proses ekstraksi informasi yang ada yang memiliki kegunaan, belum diketahui sebelumnya, dan bersifat potensial. Menurut Kusrini & Luthfi (2009) data *mining* merupakan proses automatis terhadap data yang sangat besar yang sudah ada untuk mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat. "Dalam penelitian Turban *et al.* data *mining* adalah proses yang menggunakan teknik statistik, matematika, keceradan buatan dan machine learning untuk mengekstraksi dan mengidentifikasi informasi pengetahuan yang potensial dan bermanfaat yang tersimpan dalam database besar" (Kusrini & Luthfi, 2009). Dari beberapa uraian definisi yang sudah disampaikan data *mining* dapat dituliskan sebagai *proses ekstraksi dari suatu informasi pengetahuan atau pola yang sifatnya penting dan bermanfaat dalam basis data yang memiliki rukuran besar dan* merupakan bagian integral dari KDD. Salah satu teknik dalam data *mining* adalah *clustering*.

Clustering merupakan pengelompokkan objek menjadi *cluster-cluster* yang memiliki similaritas tinggi antar objek dalam satu *cluster* tetapi tidak dengan objek *cluster* lain (Han *et al*., 2011). Similaritas atau kemiripan dan ketidakmiripan dapat ditentukan melalui nilai atribut yang mendiskripsikan objek. Menurut Abu & Supriyono (2004), *clustering* adalah teknik yang dapat digunakan dalam mempartisi serangkaian data menjadi beberapa group berdasarkan similaritas yang telah ditentukan sebelumnya. Selain itu, *clustering* juga dapat diartikan sebagai proses partisi atau pengelompokan dari satu set objek data menjadi himpunan bagian atau *cluster*. Dari beberapa definisi yang sudah disampaikan, *clustering* merupakan salah satu bagian dari data *mining* untuk mengumpulkan objek dari cluster berdasarkan kesamaan-kesamaan yang dinilai berdasarkan atribut yang mendeskripsikan objek sehingga membentuk beberapa grup. Salah satu data yang dapat ditentukan kelas *cluster*-nya dengan menggunakan algoritme *clustering* adalah data kategorik.

1.1. Data Kategorik

Menurut Kantardzic (2011), data kategorik merupakan data simbolik berupa variabel yang memiliki dua relasi. Pada umumnya data kategorik merupakan data dari hasil observasi. Variabel dalam data kategorik berupa variabel kategori. Biasanya data kategorik digunakan dalam tipe data yang tidak bisa dihitung secara kuantitatif atau perhitungan secara matematis seperti perkalian, pengurangan, pembagian, atau penjumlahan. Dalam kutipan Agresti (1990) variabel kategori adalah variabel yang skala pengukurannya terdiri dari sekumpulan kategori (Alamsyah, 2006). Data kategorik dapat dituliskan sebagai data hasil pengamatan yang bersifat *non-numeric* yang memiliki variabel dengan skala pengukuran dari sekumpulan kategori. *Clustering* data kategorik dilakukan dengan menggunakan ukuran kemiripan atau similaritas.

1.2. Ukuran Kemiripan (Similaritas)

Menurut Chatfield & Collins (2013) ukuran kemiripan dan ketakmiripan pada dasarnya merupakan kebalikan dalam arti berlawanan arah. Ukuran kemiripan yang juga sering disebut koefisian kemiripan yang biasanya bernilai antara 0 dan 1.

Menurut Guha (1999) $\text{sim}\left(T_i, T_j\right) = \left|\frac{T_i \cap T_j}{T_i \cup T_j}\right|$. Sim $\left(T_i, T_j\right)$ merupakan besarnya nilai kemiripan $0 \leq \theta \leq 1$ antara objek T_i dan T_j dengan $\left|T_i \cap T_j\right|$adalah banyaknya irisan variabel yang sama antara objek T_i dan T_j dan $\left|T_i \cup T_j\right|$ adalah banyaknya gabungan variabel yang sama antara objek T_i dan T_j.

1.3. Algoritme Robust Clustering using linKs (ROCK)

- Menentukan inisialisasi nilai k, θ, dan anggap satu titik sebagai kelompok yang terpisah.
- Hitung *link* dari setiap pasang kelompok yang merupakan *neighbours.*
- Selain similaritas, dalam beberapa algoritme *clustering* juga menggunakan *link* dalam proses algoritmenya. Menurut Guha (1999) *link* $\left(T_i, T_j\right)$ antar dua obyek didefinisikan sebagai banyaknya *neighbors* antara obyek T_i dan T_j. Secara intuisi jika *link* $\left(T_i, T_j\right)$ besar maka kemungkinan besar bahwa T_i dan T_j.masuk dalam kelompok yang sama.
- Gabungkan dua *cluster* yang mempunyai *God of Measure* terbesar sebagai satu *cluster*.
  Menurut Guha *et al.* (2000), *Goodness of Measure* $g\left(C_i, C_j\right)$ untuk menggabungkan dua kelompok C_i, C_j ditulis sebagai $g\left(C_i, C_j\right) = \frac{link\left(C_i, C_j\right)}{\left[\left(n_i+n_j\right)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}\right]}$, dengan $f(\theta) = \frac{1-\theta}{1+\theta}$ dan $link\left(C_i, C_j\right)$ merupakan jumlah *cross link* antara objek-objek dalam C_i, C_j ditulis sebagai $link\left(C_i, C_j\right) = \sum_{T_i \in C_i, T_j \in C_j} link\left(T_i, T_j\right)$.
- Kembali ke langkah menghitung *link.* Penggabuingan dilakukan sampai memenuhi salah satu kriteria, yaitu.
  - Sejumlah *cluster* tertentu (k) diperoleh.
  - Tidak ada *link* yang tersisa diantara *cluster-cluster* yang terbentuk.

1.4. Quick Robust Clustering using linKs (QROCK)

Quick Robust Clustering using linKs (QROCK) merupakan penyempurnaan dari algoritme ROCK. Algoritme ini hampir mirip dengan algoritme ROCK dari segi alurnya, akan tetapi dalam algoritme QROCK menggunakan similaritas dengan perhitungan similaritas terbobot.

QROCK menggunakan primitive tipe data abstrak MFSET untuk menggabungkan *cluster-cluster*-nya. MFSET (*Merge Find Set*) merupakan struktur dari suatu data dengan dua operasi yaitu:

- *Find* : penentuan set yang terdiri dari elemen khusus yang berfungsi untuk membentuk dua elemen pada suatu set.
- *Merge* : penggabungan dari dua set menjadi satu set.

MFSET yang ada dalam algoritme QROCK yaitu:

- *Merge* : penggabungan antara komponen A dan B.
- *Find* (*x*) : pencarian komponen yang salah satu anggotanya adalah *x*.
- *Initial* (*x*) : membentuk komponen yang memuat elemen *x*.

1.5. Algoritme Quick Robust Clustering using linKs (QROCK)

Menurut Dutta *et al*. (2005) algoritme QROCK diuraikan sebagai berikut:

- Menentukan inisialisasi dari data poin sebagai *cluster* awal.
  Langkah awal dalam algoritme QROCK adalah melakukan inisialisasi dengan menentukan nilai θ atau besarnya nilai similaritas awal.
- Menghitung nilai kemiripan atau similaritas terbobot menggunakan persamaan:

$$\text{sim}\left(T_i, T_j\right) = \frac{\left|T_i \cap T_j\right|}{\left|T_i \cap T_j\right| + 2\sum_{k \notin T_i, T_j} \frac{1}{|D_k|}}$$

  $\left|T_i, T_j\right|$: banyaknya kategori yang sama antara objek T_i dan objek T_j
  $|D_k|$: selisih perbedaan tingkat kategori.
  Untuk objek yang sama maka nilai similaritasnya adalah 1.

- Mencari *neighbors* dari masing-masing objek.
  Suatu objek T_i dan T_j bertetangga jika $sim(T_i, T_j) \geq \theta$. Sehingga dapat dituliskan *neighbors* dari masing-masing objek adalah
  $neighbors\ [T_1] = [T_i, T_1]$
  $neighbors\ [T_2] = [T_i, T_2]$
  $neighbors\ [T_3] = [T_i, T_3]$
  $\vdots$
  $neighbors\ [T_j] = [T_i, T_j]$
- Melakukan *clustering*.
- Inisialisasi (i) untuk setiap objek sehingga terbentuk komponen yang memuat anggota objek i.
- Untuk setiap i mulai dari 1 sampai dengan n
  $i = 1$
  Ambil suatu nilai x dengan *neighbors* $[T_i] = [T_i, T_j] \Longrightarrow x = T_i$
  Ambil setiap nilai $y = x$ dengan *neighbors* $[T_i] = [T_i, T_j] \Longrightarrow y = T_j$
  $A = find\ (x) = find\ (T_i) = \{T_i\}$
  $B = find\ (y) = find\ (T_j) = \{T_j\}$
  Jika $A \neq B$ maka *merge* (A, B): $\{T_1, T_j\}$ sehingga terbentuk sekumpulan komponen $\{T_i, T_j\}, \cdots, \{T_j\}$.
  Jika $A = B$ maka tidak ada penggabungan komponen.
- Kembali ke langkah perhitungan i sampai dengan n, jika $i = n$ maka iterasi stop. Proses *clustering* selesai.

1.6. Aspek Akurasi Algoritme QROCK

Dalam proses *clustering* jika terdapat objek tunggal sebagai kelompok sendiri maka ada 2 kemungkinan yaitu objek tersebut sebagai outlier dari data yang tidak dapat bergabung pada *cluster* besar atau objek tersebut belum sempat bergabung karena kriteria dari proses *clustering* sudah terpenuhi. Menurut Alamsyah (2006) proses *clustering* tidak akan berhenti oleh jumlah *cluster* yang diinginkan akan tetapi proses *clustering* berhenti jika tidak ada *link* yang tersisa diantara *cluster-cluster* akhir yang terbentuk. Algoritme QROCK tidak menggunakan nilai k (jumlah *cluster* akhir) yang harus dipenuhi, sehingga dalam prosesnya algoritme QROCK akan berhenti jika tidak ada *link* yang tersisa. Dengan demikian, hasil *clustering* algoritme QROCK tidak menimbulkan objek tunggal sebagai kelompok sendiri, sehingga dapat dikatakan bahwa proses *clustering* algoritme QROCK memenuhi sifat robust (dapat mendeteksi adanya outlier).

Berikut disajikan tabel perbandingan akgoritme ROCK dan algoritme QROCK.

Tabel 1. Perbandingan algoritme ROCK dan QROCK

Algoritme	ROCK	QROCK
Melakukan inisialisasi	✓	✓
menghitung *neighbours*	✓	✓
Menghitung similaritas		✓
Menghitung *link*	✓	
Menghitung *Goodness of Measure*	✓	
Melakukan *clustering*	✓	✓

Dari tabel 1, dapat disimpulkan algoritme QROCK lebih efisien karena terdapat pemotongan langkah yaitu perhitungan *linKs* dan *Godness of Measure*.

3. Simpulan

Berdasarkan pembahasan diperoleh simpulan bahwa algoritme QROCK memenuhi sifat robust sehingga mampu mendeteksi adanya *outlier* dalam data kategorik. Adanya pemotongan langkah dalam algoritme QROCK yaitu perhitungan *linKs* dan *Goodness of Measure* membuat waktu iterasi algoritme QROCK lebih efisien.

Daftar Pustaka

Abu, A., & Supriyono, W. (2004). *Psikologi belajar*. Bandung: Pustaka Setia.

Alamsyah, M. (2006). Pengelompokan Data Kategorik Dengan Algoritma QROCK (*Doctoral Dissertation*). Universitas Airlangga.

Agresti, A. (1990). *Categorical Data Analysis*. John Willey & Sons Inc. New York.

Ariawan, P. A. (2019). Optimasi Pengelompokkan Data Pada Metode K-Means dengan Analisis Outlier. *Jurnal Nasional Teknologi dan Sistem Informasi*, *5*(2), 88-95.

Chatfield, C., & Collins, A. J (2013). Introduction to multivariate analysis. Springer.

Dewangan, R. R., Sharma, L. K., & Akasapu, A. K. (2010). Fuzzy Clustering Technique for Numerical and Categorical dataset. *International Journal on Computer Science and Engineering (IJCSE), NCICT Special Issue*, 75-80.

Dutta, M., Mahanta, A. K., & Pujari, A. K. (2005). QROCK: A quick version of the ROCK algorithm for clustering of categorical data. *Pattern Recognition Letters*, *26*(15), 2364-2373.

Goharian, N., & Grossman, D. (2003). Introduction to Data Mining. (*Online*). (http://www.ir.iit.edu/~nazli/cs422/CS422-Slides/DMintroduction.pdf, diakses 8 Oktober 2020).

Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information systems*, *25*(5), 345-366.

Guha, S. (1999). ROCK: A clustering algorithm for categorical attributes. In *15th Int. Conf. on IEEE Data Engineering*. Sydney, Australia, 1999.

Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, 83-124.

He, Z., Xu, X., & Deng, S. (2005). Clustering mixed numeric and categorical data: A cluster ensemble approach. arXiv preprint cs/0509011.

Johnson, R. A., & Wichern, D. W. (2002). Applied multivariate statistical analysis (Vol. 5, No. 8). Upper Saddle River, NJ: Prentice hall.

Kantardzic, M. (2011). Data Mining: Conceptss, Models, methods and Algorithms. John Wiley & Sons.

Kusrini, E. T. E., & Luthfi. (2009). *Algoritma Data Mining*.

Mujiasih, S. (2011). Pemanfatan Data Mining Untuk Prakiraan Cuaca. *Jurnal Meteorologi dan Geofisika*, *12*(2).