PRISMA 5 (2022): 716-720
PRISMA, PROSIDING SEMINAR NASIONAL MATEMATIKA
https://journal.unnes.ac.id/sju/index.php/prisma/
ISSN 2613-9189

Clustering Data Campuran Numerik dan Kategorik Menggunakan Algoritme *Ensemble Quick RObust Clustering using linKs* (QROCK)

Hanin Nabila Putri[a,*], Dewi Retno Sari Saputro[b]

[a,b]*Program Studi Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sebelas Maret, Jl. Ir Sutami No. 36A, Surakarta 57126, Indonesia*

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

Abstrak

Pengekstrakan informasibaru dan berguna dari basis data yang besar untuk membantu mengambil keputusan (*Knowledge Discovery*) disebut data *mining*. *Clustering* merupakan teknik data *mining* untuk melakukan pengelompokkan data yang memiliki similaritas tinggi ke dalam *cluster* yang sama. Umumnya proses *clustering* hanya untuk data numerik atau kategorik saja, namun kerap kali ditemui kasus data campuran numerik dan kategorik. Algoritme *Cluster Ensemble Based Mixed Data* (algCEBMD) adalah salah satu algoritme untuk mengoperasikan data campuran. Tahap yang dilakukan pada algCEBMD adalah mengelompokkan masing-masing data numerik dan kategorik dengan algoritme yang sesuai, kemudian hasil masing-masing algoritme digabung dan dikelompokkan dengan algoritme data kategorik. Pada penelitian ini digunakan algoritme *ensemble* QROCK dimanapengelompokkan data numerik menggunakan algoritme K-Medoid danpengelompokkan data kategorikmenggunakan algoritme *Quick RObust Clustering using linKs* (QROCK), selanjutnya hasil masing-masing kelompok digabung dan dilakukan pengelompokkan menggunakan algoritme QROCK. Pada sekumpulan data sering terdapat nilai yang terpaut jauh dari nilai umumnya atau karakteristik data tersebut sangat berbeda dengan data lainnya, yang disebut *outlier*. Tujuan penelitian ini untuk mengkaji algoritme *ensemble* QROCK terhadap data dengan*outlier*. Hasil penelitian menunjukkan bahwa algoritme K-Medoid dan QROCK memiliki sifat robust yang baiksehingga algoritme *ensemble* QROCK juga memiliki sifat robust yang baik untuk data *outlier*.

Kata kunci:

clustering, data campuran,*ensemble*, K-Medoid, QROCK, *outlier*.

1. Pendahuluan

Proses untuk memperoleh informasi yang bermanfaat dari penyimpanan basis data besar, yang juga didefiniskansebagai penggalian informasi baru dari gudang basis data besar untuk pengambilan keputusan atau *Knowledge Discovery*disebut data *mining* (Tan *et al*., 2006). Menurut Thuraisingham (2000), ada berbagai teknik dalam data *mining* dimana masing-masing teknik akan memperoleh hasil yang berbeda, yaitu klasifikasi, asosiasi, *clustering*, prediksi, estimasi, dan analisis deviasi.

Clustering merupakan sekumpulan objek data yang dipartisi menjadi himpunan bagian (subset) yang disebut *cluster*, dalam satu *cluster*objek-objek memiliki kesamaan karakteristik danberbeda dari*cluster* yang lain (Irwansyah & Faisal, 2015). *Clustering* termasuk dalam salah satu teknik analisis multivariat. Menurut Johnson & Wichern (2002), analisis yangbanyaknya variabel pada pengamatan lebih dari dua dan dianalisis secara bersama adalah analisis multivariat. Umumnya, proses algoritme *clustering*hanya diaplikasikanpada salah satu dari data kategorik atau data numerik saja. Namun, kerap kali data terdiri dari numerik dan kategorik atau disebut data campuran.

Clustering data campuran dapat menggunakan cara transformasi data numerik menjadi kategorikatau data kategorik menjadi numerik. Pada tahun 2010,Dewangan *et al*. menelitivariabel kategorik yangditransformasi menjadi variabel numerik, selanjutnya dilakukan *clustering* menggunakan algoritme

To cite this article:
Putri, H. N. & Saputro, D. R. S. (2022). *Clustering* Data Campuran Numerik dan Kategorik Menggunakan Algoritme *Ensemble Quick RObust Clustering using linKs* (QROCK).*PRISMA, Prosiding Seminar Nasional Matematika 5*, 716-720

data numerik. Namun, cara transformasi mempunyai kelemahandalammenetapkan transformasi agar informasi penting dari data aslitidak banyak yang hilang.

Selain *clustering* data campuranmenggunakan cara transformasi, terdapat algoritme yang dikembangkan oleh He *et al*. (2005)yaitu algoritme *Cluster Ensemble Based Mixed Data* (algCEBMD). *Cluster ensemble* merupakan teknik pengelompokkan dengan menggabungkan beberapa algoritme untuk mendapatkan kelompok yang robust dan lebih baik(Yoon *et al*., 2006). Tahap yang dilakukan dalam algoritme ini adalah mengelompokkan masing-masing data numerik dan kategorik menggunakan algoritme yang sesuai secara terpisah, hasil pengelompokkan digabung menggunakan *cluster ensemble. Cluster ensemble* merupakan proses penggabungan hasil dari beberapa algoritme *clustering* yang berbeda sehingga diperoleh hasil gabungan sebagai hasil akhir.

Data numerik dapat dikelompokkan menggunakan metode *partitional clustering*. Salah satu algoritme pada metode *partitional clustering*adalahK-Medoidyang merupakan pengembangan dari algoritme K-Means. Pada algoritme ini digunakan nilai median atau medoid sebagai pusat *cluster*, sehingga relatif lebih kuat terhadap *outlier* pada data K-Means. Algoritme *RObust Clustering usink linKs (*ROCK) yang dikembangkan Guha *et al*. (1999)adalah salah satu algoritme untuk data kategorik. Pada tahun 2005, Dutta *et al*. (2005)meneliti percepatan dari algoritme ROCK yaitu algoritme *Quick RObust Clustering using linKs* (QROCK). Dalam penelitiannya, algoritme QROCK dinilai lebih efisienjika dibandingkan dengan algoritme ROCK karena terdapat pemotongan langkah perhitungan *link* dan *Goodness of Measure* sehingga diperlukanwaktu iterasi yang lebih sedikit. Selain itu, algoritme QROCK juga mampu mendeteksi adanya *outlier*pada data. Oleh sebab itu, dikaji algoritme *ensemble* QROCK terhadap *outlier*pada penelitian ini.

2. Pembahasan

Data merupakan sekumpulan fakta yang diolah dan kemudian dapat menghasilkan suatu informasi. Menurut jenis variabelnya, data dibagi menjadi data numerik dan data kategorik (Anderson & Sclove, 1974). Data numerik adalah datakuantitatifyang nilainya dalam bentuk numerik (angka). Data kategorik merupakan kumpulan kategori dan setiap nilai mewakili beberapa kategori, data kategorik disebut juga data kualitatif yang berbentuk tidak beraturan (Bhagat *et al*., 2013).

Clustering adalah teknik dalam data *mining*yang berfungsi mengelompokkan sekumpulan objek ke dalam beberapa *cluster*yang memiliki kemiripan karakteristik sedemikian sehingga objek pada satu*cluster*mirip namun tidak mirip dengan objek di*cluster* yang berbeda (Han *et al*., 2012). Menurut Madhulatha (2012), *cluster* adalah kumpulan objek yang serupa diantara kelompoknya dan berbeda dengan objek milik kelompok lain.Terdapat dua metode *clustering*, yaitu *hierarchical clustering* dan *partitional clustering*. Data dikelompokkan melalui bagan hierarki pada metode *hierarchical clustering*, dimana dua grup terdekat digabung atau seluruh data dibagi ke dalam *cluster*.Pada *partitional clustering* data dikelompokkan tanpa ada struktur hierarki, setiap *cluster*mempunyai *centroid* dengan tujuan untuk meminimumkan jarak dari seluruh data ke *centroid*(Wanto *et al*., 2020).*Clustering*digunakan dalammengoperasikan data numerik dan data kategorik.

2.1. Clustering Data Numerik

Hierarchical clustering dan *partitional clustering* digunakan untuk *clustering* data numerik. Metode *hierarchical clustering* diklasifikasikan menjadi *agglomerative* dan *divisive*. Klasifikasi ini bergantung pada apakah hierarki terbentuk dari atas menuju bawah atau sebaliknya dari bawah menuju atas. Hasil dari metode ini dapat ditampilkan secara grafis sebagai pohon, yang disebut dendogram (Rani & Rohil, 2013). *Agglomerative hierarchical clustering*adalah metode pengelompokkan hierarki dari bawah ke atas, pengelompokan dimulai dengansetiap objek dalam *cluster* yang terpisah, kemudian berdasarkan jarak objek digabung sebagai satu kelompok yang besar. Pada *partitional clustering* sejumlah titik dipilih sebagai *centroid*, kemudian mengukur jarak setiap data ke setiap *centroid*, untuk data yang memiliki jarak terdekat ke suatu *centroid* tertentu maka data tersebut dikelompokkan. Saat proses pengelompokkan,*centroid* selalu diperbarui dan data dipindahkan ke *cluster* baru sampai mencapai jumlah *cluster* yang diinginkan (Salem *et al*., 2018).

Algoritme yang dapat diterapkan untuk *clustering* data numerik adalah algoritme *agglomerative* atau biasa disebut algoritme AGNES. Terdapat beberapa macam algoritme AGNES, yaitu *average linkage*, *single linkage*, dan *complete linkage*. Namun, pada algoritme AGNES memiliki beberapa kelemahan yaitu sering terjadi kesalahan pada data *outlier*, terdapat variabel yang tidak relevan, dan perbedaan ukuran jarak yang digunakan. Untuk mengatasi kelemahan tersebut dapat digunakan salah satu algoritme

pada *partitional clustering* yang juga dapat digunakan dalam *clustering* data numerik yaitu algoritme K-Means. Pusat *cluster* algoritme K-Means menggunakan nilai mean sehingga menyebabkan algoritme ini sensitif terhadap *outlier* (Han et al., 2012). Dengan demikian dikembangkan algoritme K-Medoid dimana pada algoritme ini digunakan nilai median atau medoid sebagai pusat *cluster* sehingga relatif lebih kuat terhadap *outlier* pada data(Saket J & Pandya, 2016). Selain data numerik, *clustering* juga digunakan untuk data kategorik.

2.2. Clustering Data Kategorik

Clustering dengan metode*hierarchical clustering* dan *partitional clustering* tidak tepatjika digunakan untuk data kategorik karena menggunakan jarak antara titik, sehingga dikembangkan algoritme ROCK. Algoritme ROCK mendasar pada konsep *neighbour* yang merupakan sepasang kelompok yang berada dalam ambang batas (*threshold)* bernilai antara 0 sampai 1 dan *link* antarpasangan kelompok yang dapat diartikan sebagai banyaknya *neighbour* bersama untuk sepasang kelompok. Algoritme ROCK dapat menangani *outlier* tetapi kurang efektif karena digunakan penentuan awal jumlah *cluster* yang ingin dibentuk. Penelitian Dutta *et al.* (2005)menghasilkan algoritme percepatan dari ROCK yaitu algoritme QROCK. Algoritme QROCK memotong langkah perhitungan *link* dan *Goodness of Measure* pada algoritme ROCK sehingga waktu iterasi lebih efisien. Pada algoritme QROCK juga tidak menentukan jumlah *cluster* akhir yang harus dipenuhi sehingga proses akan berhenti jika sudah tidak ada data yang tersisa dan tidak akan menimbulkan *cluster* tunggal pada hasil akhir *clustering*,dimana hasil tunggal tersebut kemungkinan merupakan *outlier* atau data yang belum bergabung karena kriteria *clustering* sudah dipenuhi. Oleh karena itu, dinilai algoritme QROCK dapat mendeteksi adanya *outlier* secara efektif (Sari & Saputro, 2021).*Clustering* tidak hanya digunakan untuk data numerik atau kategorik saja, tetapi juga untuk data campuran numerik dan kategorik. Salah satu algoritme *clustering* data campuran yaitu algoritme *Cluster Ensemble Based Mixed Data* (algCEBMD).

2.3. Algoritme Cluster Ensemble Based Mixed Data (algCEBMD)

AlgCEBMD digunakan dalam pengelompokkan data campuran numerik dan numerik. Pada algoritme ini terdiri dari dua tahap. Pertama, *clustering* data menggunakan beberapa algoritme yang sesuai dan menyimpan hasil tersebut berupa data kategorik. Kedua, menentukan hasil akhir dengan *cluster ensemble* dari hasil masing-masing kelompok pada tahap pertama yang diperolehmenggunakan algoritme *clustering* data kategorik.

He *et al*. (2005) menguraikan algCEBMD yang ditulis sebagai berikut:

- Memisahkan data menjadi data numerik dankategorik.
- Melakukan *clustering* data numerik menggunakan algoritme *clustering* yang sesuai.
- Melakukan *clustering* data kategorik menggunakan algoritme *clustering* yang sesuai.
- Menggabungkan hasil *clustering* data numerik dan kategorik.
- Melakukan *cluster ensemble*dengan algoritme *clustering* data kategorik sehingga mendapat *cluster* akhir.

2.4. Algoritme Ensemble Quick RObust Clustering using linKs (QROCK)

Algortime *ensemble* QROCK menggunakan konsep pada algCEBMD yaitu memerlukan algoritme *clustering* untuk data numerik dan kategorik. Algoritme K-Medoid untuk *clustering* data numerik dan algoritme QROCK untuk *clustering* data kategorik memiliki hasil keakuratan yang baik serta efisien, maka pada penelitian ini digunakan kedua algoritme tersebut.

Berikut algoritme *ensemble* QROCK:

- Memisahkan data menjadi data numerik dan kategorik.
- Melakukan *clustering* data numerik. Pada penelitian ini digunakan algoritme K-Medoid sebagai berikut(Nahdliyah *et al*., 2019):
  a. Menentukan jumlah *cluster* k.
  b. Menentukan pusat *cluster* (medoid) sebanyak $k$secara acak.
  c. Menghitung jarak setiap objekterhadap medoid pada tiap *cluster*lalu menempatkan objek ke medoid terdekat, setelah itu menghitung total jaraknya.
  d. Menentukan objek masing-masing *cluster*menjadicalon medoid baru secara acak.
  e. Menghitung jarak tiap objek dengan medoid baru pada tiap *cluster*lalu menempatkan objek ke calon medoid terdekat, setelah itumenghitung total jaraknya.
  f. Menghitung toal simpangan (S)dengan total jarak calon medoid baru dikurangi total jarakmedoid lama.Jika $S < 0$ maka calon medoid baru menjadi medoid baru.

g. Mengulangi langkah (d) sampai (f) hingga S > 0 atau tidak ada perubahan medoid.

- Melakukan *clustering* data kategorik. Pada penelitian ini digunakan algoritme QROCK sebagai berikut (Dutta *et al*., 2005):
  a. Menentukan inisialisasi untuk masing-masing data sebagai *cluster* awal dengan menentukan nilai θyaitu besar nilai similaritas awal.
  b. Menghitung similaritas terbobot antara *cluster* satu dan lainnya menggunakan persamaan

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

dengan
$|T_i \cap T_j|$: banyaknya kategori yang sama antara T_i dan T_j
$|D_k|$: selisih perbedaan kategori.

  c. Mencari *neighbours* dari masing-masing objek.T_i dan T_j dikatakan bertetangga jika $sim(T_i, T_j) \geq \theta$.
  d. Melakukan *clustering*.
    i. Menentukan inisialisasi (i) untuk setiap objek.
    ii. Dimulai dari 1 sampai n untuk setiap i,

$$i = 1$$

mengambil nilai x dengan *neighbours* $[T_i] = [T_i, T_j]$

$$x = T_i$$

mengambil nilai $y = x$ dengan *neighbours* $[T_i] = [T_i, T_j]$

$$y = T_j$$

$$A = find(x) = find(T_i) = T_i$$
$$B = find(y) = find(T_j) = T_j$$

$A \neq B \implies merge(A, B) = T_i, T_j$,
$A = B \implies$ tidak ada penggabungan.

    iii. Mengulangi langkah (ii) hingga jika $i = n$ maka iterasi selesai.

- Melakukan *cluster ensemble*yaitu menggabungkan hasil dari masing-masing algoritme K-Medoid dan QROCK, kemudian dikelompokkan lagimenggunakan algoritme QROCKuntuk mendapat *cluster* akhir dari data campuran numerik dan kategorik.

2.5. Algoritme Ensemble QROCK terhadap Outlier

Pada data sering terdapat *outlier*atau pencilan yaitu kelainan, penyimpangan, atau anomali. Data diolah dengan satu atau lebih proses, saat prosessering terjadi perilaku yang tidak biasa, hal itu dapat menghasilkan *outlier*. Oleh sebab itu, *outlier*sering kali berisi informasi yang berguna tentang karakteristik data sehingga tidak dapat diabaikan (Aggarwal, 2016). Algoritme *ensemble* QROCK menggunakan algoritme K-Medoid dan QROCK yang secara berurutanuntuk mengelompokkan data numerik dan kategorik.Sifat robust yang dimiliki algoritme K-Medoid terhadap *outlier* adalah baik, karenaalgoritme ini menggunakan nilai median sebagai pusat *cluster*. Algoritme QROCK juga relatif kuat terhadap *outlier* karena proses iterasi akan berhenti jika sudah tidak ada data yang tersisa sehingga tidak akan menimbulkan *cluster* tunggal pada hasil akhir. Oleh karena algoritme K-Medoid dan QROCK memiliki ketahanan terhadap data yang memiliki *outlier*, maka dapat dinilai algoritme *ensemble* QROCK untuk *clustering* data campuran numerik dan kategorik juga memiliki ketahanan yang baik terhadap data dengan *outlier*.

3. Simpulan

Berdasarkan pembahasan diperoleh simpulan bahwa algoritme *ensemble* QROCK merupakan algoritme *clustering*untuk data campuran numerik dan kategorik. Pada algoritme ini digunakanalgoritme K-Medoid untuk *clustering* data numerik dan algoritme QROCK untuk *clustering* data kategorik. Kedua algoritme memiliki sifat robust yang baik sehingga algoritme *ensemble* QROCK juga memiliki sifat robust yang baikuntuk data yang didalamnya terdapat*outlier*. Algoritme *ensemble* QROCK pun memiliki langkah-langkah yang efektif dan efisien.

Daftar Pustaka

Aggarwal, C. C. (2016). An Introduction to Outlier Analysis. In *Outlier Analysis* (hal. 1–34).

Anderson, T. W., & Sclove, S. L. (1974). *Introductory Statistical Analysis*. Houghton Mifflin.

Bhagat, P. M., Halgaonkar, P. S., & Wadhai, V. M. (2013). Review of Clustering Algorithm for Categorical Data. *International Journal of Engineering and Advanced Technology*, *3*(2), 341–345.

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)*, 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.

Guha, S., Rastogi, R., & Shim, K. (1999). ROCK: A Robust Clustering Algorthim for Categorical. *International Conference on Data Engineering*, 512–521.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques Third Edition*. The Morgan Kaufmann Series in Data Management Systems.

He, Z., Xu, X., & Deng, S. (2005). A Cluster Ensemble Method for Clustering Categorical Data. *Information Fusion*, *6*(2), 143–151.

Irwansyah, E., & Faisal, M. (2015). *Advanced Clustering: Teori dan Aplikasi*. DeePublish.

Johnson, R. A., & Wichern, D. W. and others. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall.

Madhulatha, T. S. (2012). An Overview of Clustering Methods. *IOSR Journal of Engineering*, *2*(4), 719–725.

Nahdliyah, M. A., Widiharih, T., & Prahutama, A. (2019). Metode K-Medoids Clustering dengan Validasi Silhouette Index dan C-Index. *Jurnal Gaussian*, *8*(2), 161–170.

Rani, Y., & Rohil, H. (2013). A Study of Hierarchical Clustering Algorithm. *International Journal of Information and Computation Technology*, *3*(11), 1225–1232.

Saket J, S., & Pandya, S. (2016). An Overview of Partitioning Algorithms in Clustering Techniques. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *5*(6), 1943–1946.

Salem, S. Ben, Naouali, S., & Chtourou, Z. (2018). A fast and effective partitional clustering algorithm for large categorical datasets using a k-means based approach. *Computers and Electrical Engineering*, *68*, 463–483.

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.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson Education.

Thuraisingham, B. (2000). A Primer for Understanding and Applying Data Mining. *IT Professional*, *2*(1), 28–31.

Wanto, A., Siregar, M. N. H., Windarto, A. P., Hartama, D., Ginantra, N. L. W. S. R., Napitupulu, D., Negara, E. S., Dewi, M. R. L. S. V., & Prianto, C. (2020). *Data Mining : Algoritma dan Implementasi*. Yayasan Kita Menulis.

Yoon, H. S., Ahn, S. Y., Lee, S. H., Cho, S. B., & Kim, J. H. (2006). Heterogeneous Clustering Ensemble Method For Combining Different Cluster Results. *International Workshop on Data Mining for Biomedical Applications*, 82–92.

.