PRISMA 6 (2023): 702-706
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 *K-Prototype*

Rika Wijayati[a], Dewi Retno Sari Saputro[b]

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

**rikawijayati40@student.uns.ac.id, dewiretnoss@staff.uns.ac.id*

Abstrak

Terdapat berbagai teknik dalam data *mining* dimana setiap teknik memiliki hasil yang berbeda, yaitu klasifikasi, asosiasi, *clustering*, prediksi, estimasi dan analisis deviasi. *Clustering* merupakan proses yang digunakan untuk mengelompokkan objek dalam dataset sehingga objek yang mempunyai kemiripan ditempatkan pada *cluster* yang sama. Ukuran kemiripan antar objek satu dengan objek lainnya diketahui dari ukuran jarak. Semakin kecil ukuran jarak, semakin kecil perbedaan antar objek satu dengan objek lainnya. Metode *elbow* merupakan metode untuk menentukan jumlah *cluster* optimal yang diperoleh dengan melihat presentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik. Pada umumnya proses *clustering* hanya untuk data numerik atau kategorik saja, akan tetapi kerap kali ditemui kasus data campuran numerik dan kategorik. Algoritme *K-Means* merupakan salah satu metode *clustering* konvensional yang efisien untuk data yang berukuran besar, tetapi tidak untuk data kategorikal. Algoritme *K-Prototype* merupakan pengembangan dari algoritme *K-Means* yang dapat menangani data yang bertipe campuran numerik dan katagorik. Tujuan dari penelitian ini untuk mengkaji algoritme *K-Prototype* dengan inisialisasi menggunakan metode *elbow* dalam menentukan jumlah *cluster*. Hasil penelitian menunjukkan bahwa metode *elbow* dapat digunakan untuk menentukan jumlah *cluster* optimal yang selanjutnya digunakan pada algoritme *K-Prototype*.

Kata kunci:

Clustering, ukuran kemiripan, metode *elbow*, algoritme *K-Prototype*

1. Pendahuluan

Menurut (Ahmad, Qamar and Qasim Afser Rizvi, 2015), data *mining* merupakan bagian dari *knowledge discovery in databases* (KDD) yang berarti bidang penemuan baru dan berpotensi berguna untuk data yang berukuran besar. Terdapat berbagai teknik dalam data mining dimana masing-masing teknik akan memberi hasil yang berbeda, yaitu klasifikasi, asosiasi, *clustering*, prediksi, estimasi, dan analisis deviasi. *Clustering* merupakan proses pengelompokkan objek ke dalam kelas-kelas atau *cluster* sehingga objek-objek dalam *cluster* memiliki suatu kemiripan yang tinggi tetapi tidak mirip dengan objek *cluster* lain (Han and Kamber, 2012). *Clustering* adalah kumpulan objek data yang dibagi menjadi subset yang disebut *cluster*, objek dalam *cluster* memiliki sifat yang sama atau mirip berbeda dengan *cluster* lainnya (Irwansyah dan faisal, 2015). Metode *clustering* yang baik akan menghasilkan *cluster* dengan *intraclass sililarity* tinggi dan *interclass similarity* rendah (Gates and Ahn, 2017).

Ukuran kemiripan antar objek satu dengan objek lainnya diketahui dari ukuran jarak. Semakin kecil ukuran jarak, semakin kecil perbedaan antar objek satu dengan objek lainnya (Nooraeni, Tinggi and Statistik, 2015). Metode analisis *cluster* membutuhkan suatu ukuran ketakmiripan (jarak) yang didefinisikan untuk setiap pasang objek yang akan dikelompokan. Jarak yang biasa digunakan dalam analisis penggerombolan diantaranya ukuran data numerik dan ukuran data kategorik. Ukuran data numerik digunakan untuk data bertipe numerik seperti jarak *euclidean*, *mahalanobis*, *manhatan* dan lain-lain.

To cite this article:
Wijayati R.,& Saputro, D. R. S. (2023). *Clustering Data Numerik dan Kategorik Menggunakan Algoritme K-Prototype*. *PRISMA, Prosiding Seminar Nasional Matematika 6*, 702-706

Sedangkan ukuran data kategorik terdapat ukuran rasio ketidak cocokkan, *gambaryan similaruty* dan lain-lain.

Metode *elbow* pada penelitian ini digunakan untuk menentukan jumlah *cluster* terbaik pada algoritme *K-Prototype*. Tujuan dari metode *elbow* adalah untuk memilih nilai k yang kecil dan masih memiliki nilai withinss yang rendah (Hartanti, 2020). Penentuan optimal jumlah cluster pada penelitian ini menggunakan salah satu metode analisis cluster yaitu metode *Elbow*, dengan memperhatikan nilai perbandingan (dari perhitungan *SSE* untuk setiap nilai *cluster*) antara jumlah cluster yang akan membentuk siku pada suatu titik, sehingga semakin besar jumlah cluster k maka nilai SSE akan semakin kecil.

Algoritme *K-Means* merupakan algoritme pengelompokan data dengan sistem partisi (Yunita, 2018). Algoritme *K-Means* biasa digunakan dalam teknik pengclusteran dan efisien dalam data berukuran besar. Terdapat dua tipe data dalam proses *clustering* yaitu data bertipe numerik dan data bertipe kategorik. Algoritme *K-Means* hanya cocok digunakan pada data bertipe numerik dan tidak efektif juka digunakan untuk data bertipe kategorik. Algoritme *K-Prototype* merupakan algoritme dalam *clustering*, yang digunakan untuk *clustering* data yang bertipe campuran yaitu numerik dan kategorik (Kumara, 2014). Algoritme tersebut tidak terlalu rumit dan dapat mengatasi data yang berukuran besar. Hal itu merupakan kelebihan algoritme K-Prototype (Nooraeni, Tinggi and Statistik, 2015).

2. Pembahasan

2.1. *Clustering*

Clustering merupakan proses yang digunakan untuk mempartisi komponen dalam dataset sehingga komponen yang mempunyai kemiripan ditempatkan pada *cluster* yang sama, tujuannya untuk menemukan komponen dalam data set secara efisien (Bhatia and Louis, 2004). Dalam *clustering*, data yang memiliki kemiripan dimasukkan ke dalam *cluster* yang sama, sebaliknya data yang tidak memiliki kemiripan dimasukkan dalam *cluster* yang berbeda (Yunita, 2018). Menurut (Xu and Lange, 2019) *clustering* memegang peran penting dalam jumlah data yang besar, tujuannya untuk mempartisi objek-objek berdasarkan ukuran kesamaan. Dari definisi-definisi tersebut, *clustering* merupakan cara dari data *mining* untuk menggabungkan objek dari kelas *cluster* berdasarkan kesamaan-kesamaan yang dinilai berdasarkan atribut yang mendeskripsikan objek sehingga membentuk beberapa grup.

2.2. *Ukuran kemiripan*

Dalam *clustering* diperlukan beberapa ukuran untuk mengetahui seberapa mirip objek-objek yang akan dikelompokkan kedalam *cluster* yang sama. Terdapat tiga ukuran yang digunakan dalam mengukur kemiripan antar objek yaitu ukuran korelasi, ukuran asosiasi dan ukuran jarak. Ukuran korelasi dapat diterapkan pada data dengan skala metrik, Kemiripan antar objek dapat dilihat dari koefisien korelasi antar pasangan objek yang diukur dengan beberapa variabel. Ukuran asosiasi dipakai untuk mengukur data berskala nonmetrik. Sedangkan ukuran jarak merupakan ukuran kemiripan yang paling sering digunakan. Diterapkan untuk data berskala metrik, jarak yang kecil menunjukkan tingginya tingkat kemiripan antar objek (Amalia dan Sumargo, 2019). Ukuran kemiripan yang akan digunakan dalam proses *clustering* adalah ukuran jarak. Dalam mengukur jarak kemiripan atribut numerik dapat digunakan ukuran jarak euclidean. Jarak *euclidean* merupakan perhitungan jarak dari dua buah titik dalam *euclidean space*. Semakin dekat jarak, maka semakin mirip suatu objek data tersebut dengan objek lainnya. Persamaan jarak *euclidean* ditulis sebagai

$$d_r(X_j, Z_i) = \sqrt{\sum_{l=1}^{p} (x_{jl} - z_{il})^2}$$

dengan x_{jl} adalah nilai pada atribut numerik l pada kelompok i, z_{il}l adalah rata-rata atau *prototype* atribut numerik ke l kelompok i, dan p adalah jumlah atribut numerik. Sedangkan ukuran kemiripan untuk data kategorik ditulis sebagai

$$d_c(X_j, Z_i) = \gamma_i \sum_{p=l+1}^{m} \delta(x_{jl}, z_{il})$$

dan *simple matching similarity measure* untuk data kategorik adalah

$$\delta(x_{jl}, z_{il}) = \begin{cases} 0 & (x_{jl} = z_{il}) \\ 1 & (x_{jl} \neq z_{il}) \end{cases}$$

dengan γ_i adalah bobot untuk atribut kategorik pada kelompok i yang nilainya merupakan nilai standar deviasi untuk atribut numerik pada masing-masing kelompok. Ketika x_{jl} adalah nilai atribut kategorik, z_{il} adalah modus atribut ke l kelompok i, dan m adalah jumlah atribut kategorik. Jadi ukuran kesamaan untuk data yang memiliki atribut numerik dan atribut kategorik ditulis sebagai

$$d(X_j, Z_i) = \left(\sum_{l=1}^{p} (x_{jl} - z_{il})^2 + \gamma_i \sum_{l=p+1}^{m} \delta(x_{jl}, z_{il}) \right)^{\frac{1}{2}}$$

2.3. *Metode Elbow*

Metode *elbow* merupakan suatu metode pada titik tertentu akan terjadi grafik penurunan secara drastis dengan sebuah lekukan yang tajam. Nilai itu kemudian menjadi nilai k atau jumlah *cluster* terbaik. Pencarian nilai k optimum dilakukan dengan melakukan perbandingan nilai *Sum of Square Error* (SEE) yang disajikan dalam bentuk grafik (Hadiyatullah and Dkk, 2019). Dalam menentukan jumlah *cluster*, berikut disajikan algoritme *elbow* berdasarkan *SSE*:

1. Inisialisasi awal nilai k
2. Menaikan nilai k sampai jumlah *cluster* yang ditentukan
3. Menghitung nilai SSE dari setiap k

$$SSE = \sum_{k=1}^{k} \sum_{x_i \in s_k} \|X_i - C_k\|_2^2$$

dengan:

k : jumlah *cluster* yang digunakan pada algoritme *K-Prototype*

X_i : nilai atribut dari data ke-i

C_k : jumlah *cluster* i pada *cluster* ke-k

4. Melakukan perhitungan *SSE* sampai k yang ditentukan.
5. Menganalisis hasil *SSE* yang nilai k-nya turun secara signifikan.
6. Menetapkan nilai *cluster* dilihat dari hasil SSE yang nilai k-nya turun secara signifikan.

2.4. *Algoritme K-Prototype*

Algoritme *K-Prototype* merupakan cara pengelompokan data-data yang bertipe campuran numerik dan katagorik. Algoritma ini dipilih karena sangat sederhana dari sisi kompleksitas algoritma dan mampu menangani data dengan ukuran yang sangat besar. *K-Prototype* adalah salah satu metode pengelompokkan yang berbasis partisi. Algoritme *K-Prototype* merupakan ekspansi antara K-*Means* dan *K-Modes* untuk mengatasi *clustering* pada data bertipe campuran numerik dan katagorik (Iriawan et al., 2018). *K-Means Cluster Analysis* merupakan salah satu metode analisis *cluster* non hirarki yang dapat digunakan untuk mempartisi objek ke dalam kelompok-kelompok berdasarkan kedekatan karakteristik, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu *cluster* yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan ke dalam *cluster* yang lain. Jika perubahan centroid pada *K-Means Cluster Analysis* menggunakan rata-rata, maka *K-Modes Cluster Analysis* pada perubahan centroidnya menggunakan modus. K-*Modes Cluster Analysis* juga sama tahapan analisisnya dengan *K-Means Cluster Analysis*. Perubahan yang mendasar

terdapat pada pengukuran kesamaan (similarity *measure) antara objek dengan centroid (Prototype)-nya. Proses algoritme K-Prototype diuraikan sebagai berikut.*

1. Menentukan k awal prototype yaitu $Z_1, Z_2, \dots, Z_k$ sebagai pusat *cluster* di masing-masing *cluster*.
2. Menghitung jarak objek pada dataset terhadap pusat *cluster* awal dengan menggunakan persamaan

$$d(X_j, Z_i) = \left(\sum_{l=1}^{m_r} (x_{jl}^r - z_{il}^r)^2 + \gamma_i \sum_{l=l+1}^{m_c} \delta(x_{jl}^c, z_{il}^c) \right)^{\frac{1}{2}}$$

3. Mengalokasikan setiap objek ke dalam *cluster* yang memiliki jark terdekat dengan pusat *cluster* awal.
4. Menentukan pusat *cluster* baru untuk masing-masing *cluster*.
5. Menghitung jarak objek pada dataset terhadap pusat *cluster* baru, dengan menggunakan persamaan seperti langkah 2.
6. Mengalokasikan setiap objek ke dalam *cluster* yang memiliki jarak terdekat dengan pusat *cluster* baru.
7. Mengulangi langkah 2 sampai 6 hingga tidak ada lgi perpindahan objek atau anggota masing-masing *cluster* tetap.

3. Simpulan

Berdasarkan pembahasan diperoleh simpulan bahwa algoritme *K-Prototype* merupakan algoritme *clustering* untuk data bertipe campuran numerik dan kategorik. Metode *elbow* digunakan untuk mencari ukuran *cluster* terbaik yang selanjutnya digunakan pada algoritme *K-Prototype*.

Daftar Pustaka

Ahmad, P., Qamar, S., dan Qasim A. R. S., *Techniques of Data Mining In Healthcare: A Review*, International Journal of Computer Applications, 120(15), 38-50, 2015.

Amalia, D., dan Sumargo, B., *Pengelompokan Pengguna Internet Dengan Metode K-Means Clustering*, Journal Statistic dan Aplikasinya, 3(1), 2620-8369, 2019

Bhatia, S. K., dan Louis, S., *Adaptive K-Means Clustering*, 2004.

Gates, A. J., and Ahn, Y. Y., *The Impact of Random Models on Clustering Similarity*, Journal of Machine Learning Research, 18, 1-28, 2017.

Han and Kamber. Data Mining: *Concepts and TechniquesI*, 2nd ed., 2006.

Hartanti, N., *Metode Elbow dan –Means Guna Menguukur Kesiapan Siswa SMK Dalam Ujian Nasional*. Jurnal Nasional Teknologi dan Sistem Informasi, 2(6), 82-89, 2020.

Iriawan, N., Fithriasari, K., Ulama, B. S. S., Suryaningtyas, W., Susanto, I., and Pravitasari, A. A., *Ensemble Fuzzy, K-Prototypes Density Peaks Clustering Mixed) pada Pengelompokan Data Pelamar Bidikmisi Sejawa-Timur Tahun 2016-2017*., Jurnal Ilmu Komputer Dan Informasi, 11(2), 67, 2018.

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

Kumara, I. N. S, *Optimasi Pusat Cluster K-Prototype Dengan Algoritme Genetika*, Majalah Ilmiah Teknik Elektro, 13(2), 201.

Nooraeni, R., Tinggi, S., dan Statistik, I., *Metode Cluster Menggunakan Kombinasi Algoritma Cluster K-Prototype Dan Algoritma Genetika Untuk Data Bertipe Campuran Cluster*, Jurnal Aplikasi Statistika Komputasi Statistik, 7(2), 17-17, 2015.

Xu, J., and Lange, K., *Power K-Means Clustering, 36th International Conference on Machine Learning, ICML 2019*, 2019-June, 11977-11991, 2019.

Yunita, F., *Enerapan Data Mining Menggunkan Algoritma K-Means Clustering Pada Penerimaan Mahasiswa Baru*, Sistemasi, 7(3), 238, 2018.