ANALISIS CLUSTERING MENGGUNAKAN ALGORITMA K-MODES

Vidya Handayani[1], Adiwijawa[2], Angelina Prima Kurniati[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Klasterisasi adalah proses mengelompokkan data ke dalam suatu kelas atau klaster, sehingga objek pada suatu klaster memiliki kemiripan yang sangat besar dengan objek lain pada klaster yang sama, tetapi sangat tidak mirip dengan objek pada klaster lain.

Salah satu algoritma yang sering digunakan untuk melakukan proses clustering data adalah algoritma k-means. K-means sangat populer dalam proses klasterisasi data karena efisiensinya dalam mengklaster data. Namun, algoritma ini hanya terbatas untuk pengelompokan pada data numerik, sedangkan pada kenyataannya di dunia nyata banyak juga data yang atributnya bernilai kategorik.

Untuk menangani masalah data kategorik, dalam Tugas Akhir ini akan dibahas sebuah algoritma bernama k-modes yang merupakan varian dari algoritma k-means. Sama halnya seperti algoritma k-means, algoritma k-modes ini menghasilkan solusi local optimum. Hal tersebut berkaitan dengan proses inisialisasi pada penentuan centroid awal klaster. Dalam tugas akhir ini dibahas mengenai metode penentuan inisialisasi awal pada algoritma k-modes yaitu, secara random, dan menggunakan metode frequency based.

Ditunjukkan dalam tugas akhir ini bahwa metode pemilihan k inisialisasi awal menggunakan metode frequency based memiliki tingkat akurasi yang lebih baik dalam mengelompokkan data dibandingkan dengan inisialisasi secara random.

Kata Kunci : Clustering, k-means, k-modes, frequency based

Abstract

Clustering is a process of grouping data into a class or cluster, so that the objects in a cluster has a very large similarity with other objects in the same cluster, but not similar to objects in other clusters.

One commonly used algorithm for data clustering process is the k-means algorithm. K-means is very popular in clustering data process because its efficiency for clustering data. However, this algorithm is limited to numerical data grouping, whereas in fact, in the real world there are many valuable attributes of categorical data.

To handle the problem of categorical data, in this Final Project will be discussed an algorithm called the k-modes which is a variant of k-means algorithm. Just as k-means algorithm, k-modes algorithm produces local optimum solution. This is related to the initialization process in determining the initial cluster centroid. This Final Project explains about the methods for determining first initialization of k-modes algorithm by randomly, and using frequency-based method.

It is shown in this Final Project that the selection method of first k initialization using frequency-based method which has better accuracy in grouping data compared with random initialization.

Keywords : Clustering, k-means, k-modes, frequency-based

BAB I
PENDAHULUAN

1.1 Latar belakang masalah

Data mining merupakan bidang keilmuan yang biasanya terintegrasi dengan bidang keilmuan lain seperti ilmu statistik, *artificial intellegence*, sistem Basis Data serta yang lainnya [4]. *Data mining* merupakan sebuah proses untuk menggali informasi yang tersembunyi atau menemukan pola yang bermanfaat dalam sekumpulan data yang besar. Untuk mendukung proses penggalian informasi ini terdapat task-task (proses) dalam data mining, yaitu : klasifikasi, *clustering*, dan asosiasi.

Tugas akhir ini membahas mengenai *clustering*. *Clustering* merupakan pendekatan dalam *data mining* yang cukup populer untuk mengelompokkan data [1,2]. *Clustering* dilakukan dengan cara mengelompokkan data ke dalam beberapa kelompok/*cluster* berdasarkan karakteristiknya, dimana data dalam satu *cluster* akan memiliki kemiripan karakteristik antara satu dan yang lainnya, dan sangat berbeda karakteristiknya dengan data pada *cluster* yang lainnya.

Terdapat banyak metoda yang telah dikembangkan untuk melakukan *clustering* pada data. K-means merupakan salah satu metode yang efisien dalam melakukan pengelompokan data[1,7]. Algoritma ini mengelompokkan data dengan 4 operasi dasar : (1) menentukan jumlah *cluster*/ nilai k sebagai inisiasi awal, (2) menghitung jarak antar objek dan rata-rata/centroid dari sebuah *cluster*, (3) mengalokasikan masing-masing data ke centroid yang terdekat, (4) menghitung kembali rata-rata setiap *cluster* dari objek-objek yang sudah dialokasikan pada *cluster* tersebut hingga jarak *intra cluster* sangat dekat [7]. Algoritma k-means dinilai sangat mudah dan cukup efisien dalam *data mining*. Namun, yang menjadi masalah adalah k-means ini hanya terbatas untuk data yang bertipe numerik, karena pengelompokan dilakukan dengan menghitung means/rata-rata dari suatu data dengan data yang lain, sedangkan untuk data kategorik tidak dapat dihitung means-nya, dalam kenyataannya di dunia nyata sering juga ditemukan data yang bertipe kategorik.

Untuk menangani permasalahan diatas, k-modes, sebuah algoritma baru yang merupakan variasi dari algoritma k-means, mencoba memberikan solusi untuk pengelompokan pada data bertipe kategorik [5,6]. Seperti yang juga terdapat pada algoritma k-means, terdapat sebuah permasalahan dalam algoritma k-modes ini yaitu saat menentukan k inisialisasi awal. Ada 2 metode pemilihan inisialisasi awal yang digunakan, yang pertama adalah nilai k diambil dari k *record* pertama pada data. Kedua, menentukan k inisialisasi dengan menggunakan *frequency based method* yang dibahas lebih detil pada Bab II.

1.2 Perumusan masalah

Berdasarkan latar belakang tersebut, masalah yang dikaji pada TA ini adalah :

1. Bagaimana cara melakukan *clustering* data dengan menggunakan algoritma k-modes pada data kategorik.
2. Bagaimana kualitas hasil *clustering* yang dihasilkan algoritma k-modes dengan menggunakan metode inisialisasi random dibandingkan dengan menggunakan metode *frequency based.*
3. Bagaimana pengaruh perubahan nilai k terhadap akurasi hasil *clustering* pada algoritma k-modes dengan inisialisasi menggunakan metode inisialisasi random dan *frequency based.*

Adapun batasan masalah tugas akhir ini adalah sebagai berikut :

1. Data yang digunakan adalah data dengan tipe kategorik yang sudah punya kelas (untuk keperluan evaluasi hasil *clustering*).
2. Evaluasi kualitas hasil *clustering* dilakukan dengan menghitung akurasi suatu data diklasterkan ke dalam kelas yang benar.

1.3 Tujuan

Tujuan pengerjaan Tugas Akhir ini berdasarkan rumusan masalah di atas adalah :

1. Mengimplementasikan metode inisialisasi random dan *frequency based* pada algoritma k-modes.
2. Menganalisis parameter k-modes (jumlah k, metode penentuan centroid awal random dan *frequency based*) terkait dengan performansi hasil clustering.

1.4 Metodologi penyelesaian masalah

Metode yang digunakan dalam penyelesaian tugas akhir ini adalah menggunakan metode studi pustaka atau studi literatur dan analisis dengan langkah kerja sebagai berikut :

1. Studi Literatur :
   a. Pencarian referensi, mencari referensi dan sumber-sumber lain yang layak yang berhubungan dengan *data mining*, *clustering,* algoritma K-modes
   b. Pendalaman materi, mempelajari dan memahami materi yang berhubungan dengan tugas akhir
2. Pengumpulan data
   Mencari data kategorik untuk keperluan analisis algoritma K-modes.
3. Implementasi perangkat lunak
   a. Analisis dan design perangkat lunak
      Melakukan analisis dan desain perangkat lunak, mengenai kebutuhan sistem serta fungsionalitas – fungsionalitas yang dibutuhkan dalam sistem.
   b. Pengkodean

Pembuatan program berdasarkan analisis dan desain program yang telah ditentukan pada tahap sebelumnya menggunakan teknik pemrograman berorientasi objek.

c. Pengujian
Menguji aplikasi yang telah dibuat.

4. Analisis hasil
Menganalisis hasil pengelompokan yang telah dilakukan, terhadap jumlah klaster, dan kedua metode pemilihan centroid awal terkait dengan performansi.

5. Pembuatan laporan Tugas Akhir
Mengambil kesimpulan dari hasil analisis dan pembuatan laporan tugas akhir.

BAB V
Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil analisis dan pengujian pada bab sebelumnya dalam tugas akhir ini, maka didapatkan kesimpulan :

1. Performansi hasil *clustering* lebih baik menggunakan metode random dari pada menggunakan metode *frequency based*, untuk meng*cluster* data yang berukuran kecil. Pada Tugas akhir ini dibuktikan saat percobaan menggunakan data lensa yang recordnya berjumlah puluhan, nilai akurasi cenderung lebih baik jika metode inisialisasi yang digunakan adalah metode inisialisasi random.
2. Berdasarkan pengujian akurasi algoritma k-modes terhadap ketiga data uji, didapatkan kesimpulan bahwa perubahan jumlah klaster/ nilai k tidak mempengaruhi nilai akurasi hasil pengelompokan. Pada tugas akhir ini dibuktikan saat pengujian dilakukan terhadap sejumlah nilai k yaitu k=2 sampai k=7, nilai akurasi tidak menunjukkan suatu pola yang berarti, melainkan berubah disetiap perubahan nilai k.
3. Penggunaan metode frequency based dalam penentuan inisialisasi mode awal klaster algoritma k-modes akan menghasilkan akurasi yang lebih baik dibandingkan dengan inisialiasi secara random pada data yang berjumlah besar(ratusan). Pada Tugas akhir ini dibuktikan pada saat data yang digunakan adalah data lensa yang berjumlah puluhan, akurasi lebih cenderung lebih baik apabila menggunakan metode inisialisasi random. Sedangkan saat menangani data tumor, penyakit kacang kedelai dan jamur, akurasi jauh lebih baik apabila menggunakan metode *frequency based.*
4. Jumlah kemunculan kategori pada setiap atribut juga berpengaruh dalam pengelompokan menggunakan inisialisasi frequency based. Jika jumlah kemunculan kategori dalam setiap atribut bernilai sama maka, memungkinkan data tidak berpindah cluster/ centroid tidak berubah, menyebabkan iterasi berhenti sebelum pengelompokan data maksimal.
5. Hasil akurasi *clustering* menggunakan algoritma k-modes ini juga dipengaruhi oleh data yang digunakan. Semakin besar jumlah *record* data serta semakin kecil kelas yang dibentuk, maka akurasi pengelompokan akan semakin baik. Akan tetapi, hal ini akan berlaku jika data tersebut juga memiliki struktur yang baik, seperti variasi nilai dalam setiap atribut dan *missing value.*

5.2 Saran

Dengan memperhatikan berbagai ide dan hambatan selama pengerjaan tugas akhir ini, penulis mengajukan beberapa saran diantaranya:

1. Pada sistem ini, proses preprocessing ditangani menggunakan tools lain yaitu weka 3.7, jadi disarankan sistem selanjutnya telah dilengkapi proses preprocessing untuk mempercepat proses pengklasteran.
2. Perlu dilakukan percobaan terhadap data yang lebih besar lagi (puluhan ribu sampai ratusan ribu), untuk mengetahui skalabilitas algoritma.
3. Disarankan untuk melakukan pengujian menggunakan algoritma kategorik lainnya, kemudian dibandingkan hasil akurasinya terhadap hasil clustering dengan data uji yang sama.

DAFTAR PUSTAKA

[1] **Anderberg, M. R.** 1973. Cluster Analysis for Applications, Academic Press.

[2] **Anil K. Jain, Richard C. Dubes.** 1988. *Algorithms for Clustering Data.* New Jersey : Prentice Hall, 1988. 0-13-022278-X.

[3] **Ball, G. H. And Hall, D. J.** 1967. *A Clustering Technique for Summarizing Multivariate Data,* Behavioral Science, 12, pp.153-155

[4] **Bounsaythip, Catherine and Rinta-Runsala, Esa.** 2001. *Overview of Data Mining for Customer Behavioral Modeling.* Finland : VTT Information Technology.

[5] **Han, Jiawei and Kamber, Micheline.** 2001. Cluster Analysis. *Data Mining: Concepts and Techniques* . San Francisco : Morgan Kaufmann, 2001.

[6] **Huang, Z.** 1997. A Fast Clustering Algorithm to Cluster Very Large Catergorical Data Set in Data Mining. In SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 1-8.

[7] **Huang, Z.** Extensions to the K-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304

[8] **MacQueen, J. B.** Some Methods for Classification and Analysis of Multivariate Observation, In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281-297.

[9] **Michael Steinbach and Levent Ertoz.** 2004. The challenges of clustering high dimensional data. *New Directions in Statistical Physics: conophysics, Bioinformatics, and Pattern Recognition*, -:273.

[10] **O. San, V. Huynh and Y. Nakamori,** An alternative extension Of the k-means algorithmfor clustering categorical data, Int. J. Appl. Math. Comput. Sci., vol. 14, no. 2, pp. 241-247, 2004

[11] **S.S. Khan, Dr.S. Kant.** 2007. *Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation*. IJCAI-07, 2784-2789.

[12] **Tan, Pang-ning, Michael Steinbach, and Vipin Kumar**. 2006. *Introduction to Data mining*. Pearson education, Inc.

[13] **Witten, Ian H. and Frank, Eibe.** 2005. Clustering. *Data Mining Practical Machine Learning Tools and Techniques.* San Francisco : Elsevier Inc., 2005.