CLUSTERING DATA KATEGORIK MENGGUNAKAN K-MODES DENGAN WEIGHTED DISSIMILARITY MEASURE

Lutfi Hidayat Ramdhani[1], Hetti Hidayati[2], Mahmud Dwi Suliiyo[3]

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

Abstrak

K-Modes merupakan pengembangan dari algoritma clustering K-means untuk menangani data kategorik di mana means diganti oleh modes. K-Modes menggunakan simple matching meassure dalam penentuan similarity dari suatu klaster. Di dalam penelitian ini akan memodifikasi algoritma k-modes dalam penentuan similarity dengan menggunakan weighted dissimilarity meassure karena nilai dari suatu atribut cukup berpengaruh dalam clustering. Pengujian akan menggunakan real world data set yang disediakan oleh UCI repository yang bertipe kategorik. Penilaian performa dilihat dari cluster purity yang dihasilkan. Dari hasil pengujian didapatkan cluster purity pada algoritma k-modes yang menggunakan weighted dissimilarity measure memiliki nilai yang lebih baik dari algoritma k-modes sebelumnya. Ini menunjukkan objek dialokasikan kepada cluster lebih akurat menggunakan algoritma k-modes yang baru.

Kata Kunci : clustering, data kategorik, weighted dissimilarity measure, cluster purity

Abstract

K-Modes is an extension from K-means clustering algorithm for handling categorical data where modes is used instead of means. Simple K-Modes use simple matching measure to decide similarity value from a cluster. This research will modified k-modes algorithm in deciding similarity value using weighted dissimilarity measure because the value from an attribute really affect a clustering process. The experiment will be tested using real world data sets obtained from the UCI data repository where the type of the data sets is categorical. The performance of the algorithm will be seen from the value of cluster purity from created clusters. From the test showed that the value of cluster purity using k-modes algorithm using weighted dissimilarity measure has the better result from the original k-modes. It shows that object is more accurate allocated in their respective cluster using the new improved k-modes.

Keywords : clustering, categorical data, weighted dissimilarity measure, cluster purity

1. Pendahuluan

1.1 Latar Belakang Masalah

Data mining merupakan satu bidang yang sedang berkembang saat ini dimana di dalamnya terdapat *database*, statistik, *machine learning*, dan area lain yang berkaitan dalam proses pengambilan informasi tersembunyi dari suatu data[3].

Clustering merupakan salah satu teknik dalam *data mining* yang bertujuan untuk mengelompokkan suatu object yang memiliki *similarity* tinggi dalam satu klaster dan yang memiliki *dissimilarity* tinggi di klaster yang berbeda. Data yang digunakan dalam *clustering* dapat berupa data numeric, nominal, categoric, dan bisa juga gabungan. Saat ini, *clustering* pada kategorikal data merupakan topik yang sedang banyak dilakukan dalam penelitian[2].

Pada clustering data numerik, pengukuran *distance measured* dapat menggunakan *euclidian distance* atau *Manhattan distance*. Sedangkan untuk data kategorik yang mengandung nilai seperti [male, female], [low, medium, high], pengukuran secara geometric distance tidak dapat dilakukan pada data kategorik.
Pada tahun 1998, Huang telah mengembangkan suatu metode yang dapat melakukan clustering pada data kategorik, yaitu K-Modes. Algoritma K-Modes dikembangkan untuk menemukan similarity antara object yang akan ditempatkan pada satu cluster. Contoh, jika suatu object t1= [x, y, y] dan t2= [x, s, a], jarak antara t1 dan t2 adalah 1 berdasarkan perhitungan menggunakan *hamming distance.* Perhitungan tipe ini tidak mempertimbangkan *implicit similarity* yang terkandung dalam suatu *categorical values*, yang akhirnya menghasilkan *intra cluster similarity yang* rendah dalam suatu cluster[2]. Maka di proposal ini akan mengajukan modifikasi algoritma K-modes sebelumnya yang menggunakan *simple matching* atau *miss matching measure* menjadi *weighted dissimilarity measure*. Dengan *weighted dissimilarity measure*, karena algoritma ini mempertimbangkan jumlah *atribut value* pada *cluster* dan pada data set maka akan dihasilkan intra cluster similarity yang lebih tinggi dan diharapkan akan mendapatkan *cluster* bentukan yang lebih akurat.

1.2 Perumusan masalah

Dari latar belakang di atas bisa disimpulkan beberapa permasalahan yang ada
1. Kebutuhan untuk menghasilkan cluster yang akurat pada data set yang besar bersifat data kategorik.
2. Lemahnya *intra cluster similarity* pada algoritma k-modes sebelumnya.
3. Meningkatkan akurasi menempatkan objek pada klaster yang tepat sehingga akan menghasilkan nilai *cluster purity* yang tinggi.

Terdapat batasan masalah yang ada dalam penelitian ini, yaitu:
1) Dataset yang digunakan adalah dataset yang diambil dari database UCI *Machine Learning Repository* yang mempunyai atribut kategorik, yaitu soybean (small) dan car evaluation. Pemakaian data set soybean dikarenakan data tersebut merupakan data set yang sering digunakan dalam pengujian

conceptual clustering algorithm[7].

2) Pengujian akan dibandingkan dengan pembandingan nilai *cluster purity* dari algoritma K-Modes sebelumnya dengan algoritma yang diajukan.
3) Perhitungan cluster purity akan menggunakan *clustering accuracy meassure* yang disarankan oleh Huang.

1.3 Tujuan

Menganalisis perubahan cluster purity yang dihasilkan oleh K-Modes yang menggunakan *weighted dissimilarity meassure* dengan K-Modes terdahulu dan diharapkan mendapatkan suatu metode yang memiliki performance baik dari sisi cluster purity dan scalability.

1.4 Metodologi penyelesaian masalah

Metodologi yang digunakan dan langkah-langkah dalam penyelesaian masalah yang telah dirumuskan di atas adalah:

1. Studi Literatur.

a. Pencarian referensi, mencari referensi dan sumber-sumber lain yang layak yang berhubungan dengan *data mining*, *clustering*, algoritma *k-modes* dengan *weighted dissimilarity meassure*.

b. Pendalaman materi, mempelajari dan memahami materi yang berhubungan dengan tugas akhir.

2. Pengumpulan data

Mencari data kategorik dari UCI dataset untuk keperluan implementasi algoritma *weighted dissimilarity meassure* K-Modes. Data boleh berupa data apapun, yang penting adalah karakteristik atribut dari data set tersebut merupakan data kategorik.

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. Implementasi (*coding*)

Pembuatan program berdasarkan analisis dan desain program yang telah ditentukan pada tahap sebelumnya.

c. Pengujian

Menguji aplikasi yang telah dibuat.

4. Analisis hasil

Menganalisis performansi (*cluster purity*) dari algoritma k-modes dengan *weighted dissimilarity meassure* terhadap algoritma k-modes terdahulu yang menggunakan *simple matching meassure*. Penelitian dianggap berhasil jika nilai cluster purity dari algoritma yang diajukan lebih bagus dari k-modes yang terdahulu.

5. Pembuatan laporan Tugas Akhir

Mengambil kesimpulan dari hasil analisis dan pembuatan laporan tugas akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil pengujian dan analisis yang telah dilakukan, dapat diambil beberapa kesimpulan sebagai berikut:

1. Khusus untuk dataset soybean(small), tidak bisa ditentukan algoritma mana yang menghasilkan lebih bagus *cluster purity* dikarenakan hasilnya tidak konsisten. Ini bisa disebabkan karena dimensi data dari soybean terlalu kecil dan pemilihan modes awal menggunakan metode random sehingga hasil yang didapatkan tidak optimal. Sedangkan untuk dataset car evaluation, terlihat jelas bahwa algoritma kmodes dengan *weighted dissimilarity measure* lebih optimal disbanding algoritma kmodes terdahulu. Dapat disimpulkan bahwa improved k-modes bagus untuk dataset yang memiliki jumlah objek yang banyak.
2. Selain k, faktor lain yang berpengaruh terhadap hasil klaster adalah pemilihan modes atau klaster awal.

5.2 Saran

Beberapa saran yang dapat diberikan antara lain:

1. Untuk penelitian selanjutnya, cari metode yang optimal dalam penentuan modus awal atau klaster awal.
2. Untuk melihat performansi algoritma, sebaiknya gunakan pada dataset yang tidak memiliki missing value terlebih dahulu.
3. Untuk parameter pengujian dapat ditambahkan parameter lain, yaitu waktu.

Referensi

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

[2] Aranganayagi, S., Tangavel, K. 2009. *Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure.*

[3] Fung, Glenn. 2001, *A Comprehensive Overview of Basic Clustering Algorithms*.

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

[5] Hariz, S.B., Elouedi, Z., Mellouli, K. 2007. *Selection Initial modes for Belief K-modes Method.*

[6] He, Zhengyou. *Farthest-Point Heuristic based Initialization Methods for K-Modes Clustering*.

[7] Huang, Z. 1998. *Clustering Categorical Data with k-modes*.

[8] Huang, Z. 1998. *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*. Data Mining and Knowledge Discovery, 1998, 2: 283-304

[9] Huang, Z. And Ng, K. Michael. 1999. *A Fuzzy k-Modes Algorithm for Clustering Categorical Data*.

[10] http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/KDD3.htm, terakhir diakses pada tanggal 15 Juni 2013

[11] Mar, O., Huynh, V. N, Nakamori, Y. 2004. *An Alternative Extension of the k-means Algorithm for Clustering Categorical Data.* Int. J. Appl. Math. Comput. Sci., 2004, Vol. 14, No. 2, 241–247

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

[13] Tsekouras, G., Papageorgiou, D., Kotsiantis, S., Kalloniatis, C., Pintelas, P. 2005. *Fuzzy Clustering of Categorical Attributes and its Use in Analyzing Cultural Data.*