ANALISIS DAN IMPLEMENTASI CLUSTERING CATEGORICAL DATA MENGGUNAKAN ALGORITMA CHAMELEON

Sheila Rifana Awwalia[1], Arie Ardiyanti Suryani[2], Shaufiah[3]

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

Abstrak

Clustering adalah proses mengelompokkan objek ke dalam suatu kelompok (cluster) sehingga objek memiliki kemiripan sangat besar dengan objek lain yang berada pada cluster yang sama, tetapi memiliki ketidakmiripan yang besar dengan objek yang berada pada cluster berbeda.

Algoritma hierarchical clustering mengelompokkan objek dengan membuat suatu hirarki dimana objek yang mirip akan ditempatkan pada hirarki yang berdekatan dan objek yang tidak mirip akan ditempatkan pada hirarki yang berjauhan. Permasalahan timbul ketika algoritma hierarchical clustering yang sudah ada seperti CURE dan ROCK hanya memperhatikan informasi tentang closeness dan interconnectivity antar pasangan cluster yang akan dilakukan merge tanpa memperhatikan internal closeness dan internal interconnectivity di dalam masing-masing cluster tersebut. Selain itu dalam melakukan merging pasangan cluster, CURE hanya memperhatikan informasi closeness tanpa memperhatikan informasi interconnectivity dan ROCK hanya memperhatikan informasi interconnectivity tanpa memperhatikan informasi closeness diantara pasangan cluster tersebut. Hal ini dapat berakibat pada kesalahan pengambilan keputusan dalam melakukan merge pasangan cluster. Untuk mengatasi masalah tersebut, dalam Tugas Akhir ini akan diterapkan algoritma CHAMELEON yang melakukan klasterisasi dengan cara melakukan merge pasangan cluster dengan memperhatikan informasi tentang relative closeness dan relative interconnectivity dengan parameter nilai k (jumlah tetangga pada k-nearest neighbor), nlevel (jumlah level partisi), θRI (threshold RI), dan θRC (threshold RC).

Ditunjukkan dalam Tugas Akhir ini bahwa algoritma CHAMELEON menghasilkan kualitas cluster yang baik untuk nilai parameter nlevel yang lebih besar. Sedangkan untuk ketiga parameter lainnya, secara umum nilai optimal untuk ketiga parameter tersebut (k, θRI, dan θRC) tidak dapat ditentukan secara pasti.

Kata Kunci : clustering, chameleon, hierarchical clustering, relative closeness, relative interconnectivity, parameter

Abstract
Clustering is the process of grouping objects into a group (cluster) so that the object has a very great similiarity with other objects that are on the same cluster, but has a great dissimiliarity with objects that are in different clusters.

Hierarchical clustering algorithms classify objects by creating a hierarchy where the object will be placed in a hierarchy similiar to the adjacent and similiar object will be placed on a distant hierarchy. Problem arise when hierarchical clustering algorithms that already exist such as CURE and ROCK considers only information about the closeness and interconnectivity without considers internal closeness and internal interconnectivity in these clusters. In addition to merging pairs in the cluster, CURE only pay attention to information regardless of the closeness and ROCK interconnectivity information only for the interconnectivity of information regardless of the closeness between the couple information clusters. This can result in the decision making errors inperforming the merge cluster pair. To overcome these problems, the final project will be implemented algorithms that perform clustering Chameleon with how to merge the cluster pair with respect to information about the relative closeness and relative interconnectivity with the parameter values of k (number of neighbors in k-nearest neighbor), nlevel (number of partition level), θRI (threshold RI), and θRC (threshold RC).

Indicated in this final project that Chameleon algorithm produce good quality clusters for parameter values greater nlevel. As for the three other parameters, in general optimal value for those parameters (k, θRI, and θRC) can not be determined exactly.

Keywords : clustering, chameleon, hierarchical clustering, relative closeness, relative interconnectivity, parameter

1. Pendahuluan

1.1 Latar belakang masalah

Clustering adalah salah satu proses dari *data mining* yang merupakan bagian dari proses KDD (*Knowledge Discovery of Data*). Tujuan dari *clustering* adalah untuk mengelompokkan objek-objek data ke dalam *cluster-cluster* berdasarkan persamaan dan perbedaan yang dibawa oleh masing-masing atribut objek. *Cluster* yang baik akan memiliki persamaan (*similiarity*) intra *cluster* yang tinggi dan perbedaan (*dissimilarity*) antar *cluster* yang tinggi [3].

Ada beberapa pendekatan yang digunakan dalam mengembangkan metode *clustering*. Dua pendekatan utama adalah *clustering* dengan pendekatan partisi dan *clustering* dengan pendekatan hirarki. *Clustering* dengan pendekatan partisi atau sering disebut dengan *partition-based clustering* mengelompokkan data dengan memilah-milah data yang dianalisa ke dalam *cluster-cluster* yang ada. *Clustering* dengan pendekatan hirarki atau sering disebut dengan *hierarchical clustering* mengelompokkan data dengan membuat suatu hirarki berupa dendogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hirarki yang berjauhan [3].

Algoritma *agglomerative hierarchical clustering* yang ada pada saat ini seperti CURE dan ROCK memiliki kekurangan utama dalam hal keputusan *merge* pasangan *cluster*. Dalam penggabungan *cluster* nya, algoritma CURE hanya mempertimbangkan *closeness* diantara representative point dari dua *cluster* dan mengabaikan informasi tentang *internal closeness* dari masing-masing *cluster*. Selain itu, CURE juga mengabaikan informasi *relative interconnectivity* dari dua *cluster* yang akan dilakukan *merge.* Sedangkan pada algoritma ROCK, hanya mempertimbangkan *interconnectivity* saja dan mengabaikan informasi tentang *internal interconnectivity* dari masing-masing *cluster* tersebut. Selain itu algoritma ROCK juga mengabaikan informasi *relative closeness* dari dua *cluster* yang akan dilakukan *merge*. Hal ini dapat berakibat pada kesalahan pengambilan keputusan dalam melakukan *merge* pasangan *cluster* [4]. Jika hal ini terjadi, maka kualitas *cluster* yang ditunjukkan pada nilai *cohesion* dan *cohesion/separation* yang dihasilkan oleh kedua algoritma tersebut akan kurang baik. Objek-objek yang berada pada *cluster* yang sama adalah objek-objek yang tidak memiliki kemiripan tinggi sehingga nilai *cohesion* yang dihasilkan adalah rendah. Oleh karena itu dengan semakin rendah nilai *cohesion* maka kualitas *cluster* yang ditunjukkan pada nilai *cohesion/separation* akan mengalami penurunan. Jika hal ini terjadi, maka kualitas *cluster* yang dihasilkan oleh kedua algoritma tersebut akan kurang bagus.

Salah satu pendekatan yang diterapkan untuk menangani masalah di atas adalah dengan menggunakan algoritma CHAMELEON. Algoritma CHAMELEON juga termasuk dalam *agglomerative hierarchical clustering.* Algoritma CHAMELEON sendiri mengelompokkan data dengan menggunakan dua buah fase yang berbeda yaitu fase *graph partitioning* dan fase *hierarchical agglomerative*. Proses *clustering* dimulai dengan memodelkan *similiarity* antar *cluster* dengan

membangun *graph* menggunakan pendekatan *K-nearest neighbor*. Fase *graph partitioning* dilakukan untuk menemukan kelompok-kelompok *cluster* yang saling terhubung. Kemudian fase *hierarchical agglomerative clustering* dilakukan untuk menggabungkan pasangan *subcluster* berdasarkan nilai *relative interconnectivity* dan *relative closeness* yang dimiliki oleh pasangan *subcluster* tersebut. Algoritma CHAMELEON hanya akan menggabungkan pasangan *subcluster* yang mempunyai nilai *relative interconnectivity* dan *relative closeness* yang nilainya diatas nilai *threshold RI* (*threshold relative interconnectivity*) dan *threshold RC* (*threshold relative closeness*) yang diinputkan oleh *user* [4].

Oleh karena itu, hipotesa awal yang diperoleh adalah hasil klasterisasi yang dibangun dengan algoritma CHAMELEON dapat menghasilkan kualitas *cluster* yang baik dimana *cluster* yang dihasilkan memiliki kesamaan karakteristik yang tinggi dalam satu *cluster* dan memiliki perbedaan yang tinggi antar *cluster*nya.

1.2 Perumusan masalah

Berdasarkan latar belakang tersebut, maka dapat dirumuskan permasalahan sebagai berikut:

1. Bagaimana pengaruh perubahan nilai *k* (jumlah tetangga terdekat pada *k nearest neighbor graph*), *threshold RI, threshold RC, nlevel* yang ditetapkan oleh user terhadap akurasi hasil *cluster* pada algoritma CHAMELEON.
2. Bagaimana kualitas hasil *cluster* yang dihasilkan algoritma CHAMELEON dengan melihat hasil evaluasi *cluster* menggunakan perhitungan *cohesion* dan *separation*.

Adapun batasan masalah tugas akhir ini adalah sebagai berikut :

1. Dataset yang digunakan adalah dataset yang ada pada database UCI Machine Learning Repository yang mempunyai atribut kategoris.
2. Evaluasi kualitas hasil *cluster* dilakukan dengan mengukur nilai *cohesion* (kemiripan objek intra *cluster)* dan *separation* (ketidakmiripan objek antar *cluster)*).

1.3 Tujuan

Tujuan dari penelitian tugas akhir ini adalah:

1. Menganalisis algoritma CHAMELEON terhadap perubahan parameter nilai *k* (jumlah tetangga terdekat pada *k nearest neighbor graph*), *threshold RI, threshold RC, nlevel* terkait dengan akurasi hasil *cluster*.
2. Menganalisis kualitas *cluster* yang dihasilkan dengan melihat nilai *cohesion* dan *separation*.

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

2. Pengumpulan data
   Mencari data kategoris untuk keperluan analisis algoritma CHAMELEON
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 hasil *cluster* yang terbentuk terkait dengan pengubahan parameter *k* (pada *k nearest neighbor*), *threshold RI, threshold RC, nlevel*
5. Pembuatan laporan Tugas Akhir
   Mengambil kesimpulan dari hasil analisis dan pembuatan laporan tugas akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan percobaan dan analisis yang dilakukan, maka dapat diambil beberapa kesimpulan sebagai berikut:

1. Penentuan nilai parameter *k* adalah sangat penting. Nilai tersebut tidak boleh terlalu kecil dan terlalu besar. Berdasarkan hasil pengujian, tren nilai *cohesion/separation* akan mengalami peningkatan seiring dengan nilai *k* yang diberikan semakin tinggi, namun akan mengalami penurunan pada saat nilai tertentu untuk nilai *k* yang lebih besar.
2. Nilai *nlevel* partisi yang semakin besar maka nilai *cohesion* rata-rata dan kualitas *cluster* yang ditunjukkan pada nilai *cohesion/separation* akan semakin baik.
3. Jumlah *cluster* dipengaruhi oleh nilai parameter θ_{RI} (*threshold RI*) dan θ_{RC} (*threshold RC).* Semakin besar nilai θ_{RI} dan θ_{RC} maka jumlah *cluster* yang dihasilkan semakin besar. Pemilihan nilai θ_{RI} dan θ_{RC} yang tepat akan menghasilkan kualitas *cluster* yang bagus.
4. Hasil pengujian dan analisa di atas menunjukkan tidak ada sebuah nilai yang pasti dari parameter $\theta_{RI,}$ dan θ_{RC} yang dapat memberikan hasil yang paling optimal pada setiap data berbeda yang telah diujikan. Untuk data *lenses* pemilihan $\theta_{RI,}$ dan θ_{RC} yang semakin besar akan menyebabkan kualitas *cluster* semakin menurun. Untuk data *car*, pemilihan $\theta_{RI,}$ dan θ_{RC} yang semakin besar akan menyebabkan kualitas *cluster* yang semakin baik. Sedangkan untuk data *congressional-voting*, tidak ada ketentuan untuk nilai $\theta_{RI,}$ dan θ_{RC} yang semakin besar akan menyebabkan kualitas *cluster* semakin baik atau buruk.

5.2 Saran

Sebagai saran untuk perkembangan Tugas Akhir selanjutnya, perlu dilakukan pertimbangan analisis sebagai berikut:

1. Pemilihan metode lain untuk melakukan partisi *graph* pada fase *graph partitioning* agar menghasilkan kualitas *cluster* yang lebih bagus.
2. Pengujian selanjutnya dapat dilakukan dengan membandingkan algoritma CHAMELEON dengan algoritma *clustering* untuk data kategoris lainnya.

Daftar Pustaka

[1] Osinkski, Stanislaw. 2003. *An algorithm for clustering of web search result*. Poznan: Poznan University of Technology.

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

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

[4] George Karypis, Eui-Hong(Sam) Han and Vipin Kumar; CHAMELEON*: A hierarchical clustering algorithm using dynamic modelling*; Proceedings of the IEEE International Conference on Data Engineering, Sydney, March 1999.

[5] Sudipto Guha, Rajeev Rastogi and Kyuseok Shim; *ROCK: A robust clustering algorithm for categorical attributes*; Proceedings of the IEEE International Conference on Data Engineering, Sydney, March 1999.

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

[7] Shyam Boriah, Varun Chandola and Vipin Kumar. Similiarity measures for categorical data-a comparative study. Technical Report 07-022, Department of Computer Science & Engineering, University of Minnesota, October 2007.

[8] B. W. Kernighan and S. Lin. An Efficient Heuristic Procedure for Partitioning Graphs. The Bell System Technical Journal, 49(2):291-307, 1970.

[9] Kusrini dan Emha Taufiq Luthfi. 2009. Algoritma data mining. Yogyakarta: Andi Offset.

[10] Tetyana Shatovska, Tetiana Safonova and Lirii Tarasov. 2008. The new software package for dynamic hierarchical cluster for circles types of shapes. Institute of Information Theories and Application FOI ITHEA.

[11] UCI Machine Learning Repository. Data lenses, Congressional voting, and Car. http://www.ics.uci.edu/mlearn/MLRepository. Diakses pada tanggal 24 Oktober 2010.

[12] Saurav Sahay. 2009. Study and Implementation of Chameleon Algorithm for Gene Clustering. Cognitive Computing Lab, School of Interactive Computing, Georgia Tech Atlanta, USA.

[13] yudiagusta.wordpress.com/clustering/. Diakses tanggal 2 November 2010.

[14] www.mis.ccu.edu.tw/user/yfyen/pdf/cluster2.ppt. Diakses tanggal 2 November 2010.