PENGELOMPOKAN DATA MENGGUNAKAN HIERARCHICAL CLUSTERING (AHC)

Novialita Pitaloka[1], Kiki Maulana[2], Angelina Prima Kurniati[3]

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

Abstrak

Data merupakan salah satu sumber yang digunakan untuk memperoleh suatu informasi. Namun tidak semua data dapat dimanfaatkan dengan baik. Jika data tersebut memiliki struktur yang kompleks, maka akan sulit untuk dimengerti. Sebagai contoh adalah data tagihan pelanggan PT.Telkom yang digunakan pada Tugas Akhir ini. Data tersebut memiliki jumlah record yang banyak dengan atribut yang banyak pula. Oleh karena itu diperlukan suatu proses pengelompokan yang bertujuan untuk membagi data tersebut ke dalam jumlah yang lebih sedikit sehingga proses penganalisisan data menjadi semakin mudah. Tugas Akhir ini mengimplementasikan salah satu teknik data mining yaitu clustering untuk melakukan pengelompokan data. Metode clustering yang digunakan adalah Agglomerative Hierarchical Clustering (AHC). Agglomerative Hierarchical Clustering adalah suatu metode hierarchical clustering yang bersifat bottom-up yaitu menggabungkan n buah klaster menjadi satu klaster tunggal. Metode ini dimulai dengan meletakkan setiap objek data sebagai sebuah klaster tersendiri (atomic cluster) dan selanjutnya menggabungkan klaster-klaster tersebut menjadi klaster yang lebih besar dan lebih besar lagi sampai akhirnya semua objek data menyatu dalam sebuah klaster tunggal. Kunci dari metode AHC adalah perhitungan proximity antara 2 klaster. Perhitungan ini terbagi menjadi 3 yaitu Single Linkage (jarak terkecil), Complete Linkage (jarak terbesar) dan Average Linkage (jarak ratarata). karena metode hirarki tidak dapat menghasilkan klaster secara langsung, maka digunakan metode cophenet distance untuk menganalisis hasil hirarki yang terbentuk. Dari hasil yang didapat menunjukkan bahwa Agglomerative Hierarchical Clustering (AHC) dapat digunakan untuk pengelompokan data.

Kata Kunci : AHC, Single Linkage, Complete Linkage, Average Linkage,

Abstract

Data is one of resources which used for gathering information. However, not all data working well. If the data have a complex structure, it is hard to understand. For example, data of customer invoice in PT.Telkom which used in this final project. This data have sum up the record is to lot of with the attributes amount which is there also many. Therefore, we need grouping process which is dividing data into slimmer amount so process the data analysing become progressively easy to. This Final Project is inplements one of technique in data mining which is clustering to do grouping data. The clustering method that is used is Agglomerative Hierarchical Clustering (AHC). Agglometarive Hierarchical Clustering is a method of hierarchical clustering having the character of bottom up which is joining n cluster become one single cluster. This method has begin with placing each data object as one separate cluster (atomic cluster) and join that cluster-cluster become ones large cluster and bigger again untuil the last all of data object one in one single cluster. The keys from AHC method is calculation proximity between 2 cluster. This calculation is divisible become 3 which single linkage (shortest distance), complete linkage (longest distance) and average linkage (average distance). Because hierarchy method cannot result the cluster directly so we used a cophenetic distance method to analyse result of formed hierarchy. From result is in can indicate that Agglomerative Hierarchical Clustering (AHC) applicable to grouping data.

Keywords : AHC, Single Linkage, Complete Linkage, Average Linkage,

BAB 1
PENDAHULUAN

1.1 Latar belakang

Data merupakan salah satu sumber yang dapat digunakan untuk memperoleh informasi. Akan tetapi, tidak jarang kumpulan data tersebut dibiarkan begitu saja seakan-akan menjadi kuburan data, sehingga diperlukan suatu metode yang dapat dipakai untuk menggali informasi sebanyak mungkin dari data tersebut. *Data Mining* sebagai salah satu ilmu di bidang teknologi informasi, dapat digunakan untuk mengekstraksi informasi berharga yang sebelumnya tidak diketahui dari suatu *database*. Sebagai contoh adalah data tagihan Pelanggan PT Telkom yang akan digunakan pada Tugas Akhir ini.

Salah satu informasi yang dapat digali dari data tersebut adalah pengelompokan pelanggan. Hal ini dilakukan untuk mendukung strategi manajemen yang bisa jadi berbeda untuk tiap kelompoknya. Data pelanggan ini terdiri dari beberapa atribut dengan jumlah *record* yang banyak sehingga diperlukan suatu proses *data mining* yang dapat mengelompokkan data tersebut, yaitu *clustering*. Dengan menggunakan *clustering* diharapkan dapat memberikan prediksi pengelompokan pelanggan tersebut.

Salah satu metode *clustering* yang dapat digunakan untuk mengelompokkan data adalah *Agglomerative Hierarchical Clustering (AHC). Agglomerative Hierarchical Clustering (AHC)* merupakan suatu pengelompokan hirarki yang bersifat *bottom up* dimana keberadaan setiap titik data dalam klaster ditentukan oleh *proximity* antar titik tersebut. Metode *Agglomerative Hierarchical Clustering (AHC)* yang akan digunakan dalam Tugas Akhir ini ialah *Single linkage* (jarak terkecil), *complete linkage* (jarak terjauh) dan *average linkage* (jarak rata-rata). Metode ini berawal dari objek-objek individual yang paling mirip dikelompokkan dan kelompok-kelompok awal ini digabungkan sesuai dengan kemiripannya, berulang hingga menjadi satu *cluster* tunggal.

Dengan metode ini, data pelanggan akan direpresentasikan ke dalam bentuk hirarki klaster yang selanjutnya akan dikelompokkan ke dalam kelompok-kelompok yang berbeda. Selain itu, akan dihitung juga *cophenetic correlation coefficient* untuk mengukur seberapa baik sebuah *hierarchical clustering* memenuhi kesesuaian data. Kemudian dilakukan analisa dari hasil pengelompokan yang menggunakan metode *single linkage, complete linkage dan average linkage* untuk mengetahui hirarki yang terbaik.

1.2 Perumusan masalah

Berdasarkan latar belakang masalah, maka permasalahan yang akan diangkat dalam Tugas Akhir ini, yaitu :

1. Bagaimana mengimplementasikan *Agglomerative Hierarchical Clustering* (AHC) dengan pendekatan *single linkage, complete linkage* dan *average linkage* untuk mengelompokkan suatu data.
2. Bagaimana menentukan jarak antar setiap titik data.
3. Bagaimana mengukur kesesuaian data hasil *hierarchical clustering* dengan metode *cophenetic distance* untuk memperoleh *hirarki* yang terbaik.

Dalam Tugas Akhir ini ada beberapa batasan masalah yaitu :

1. Data yang akan digunakan sebagai studi kasus adalah data tagihan pelanggan layanan PT.Telkom.
2. Data yang akan digunakan dalam format MS.Excel dengan tipe *.csv.

1.3 Tujuan

Secara umum tujuan yang ingin dicapai dalam Tugas Akhir ini adalah :

1. Mengimplementasikan metode *Agglomerative Hierarchical Clustering (AHC)* untuk pengelompokan data dalam sebuah perangkat lunak.
2. Menerapkan metode *data mining, Agglomerative Hierarchical Clustering (AHC)* untuk membentuk hirarki dari data .
3. Memberikan hasil pengelompokan data menggunakan metode *Agglomerative Hierarchical Clustering (AHC)* dengan pendekatan *single linkage, complete linkage* dan *average linkage* serta analisis hasil hirarkinya dengan *cophenetic distance*.

1.4 Metodologi Penyelesaian Masalah

Metode penyelesaian masalah yang dilakukan dalam Tugas Akhir ini mencakup hal-hal berikut :

1. Mencari dan mengumpulkan bahan-bahan literatur yang berhubungan dengan permasalahan ini, meliputi : *Data Mining, Clustering, Agglometarive Hierarchical Clustering (AHC), single lingkage, complete lingkage, average lingkage, cophenetic distance* dan pengukuturan evaluasi.
2. Studi literature tentang *Data Mining, Clustering, Agglometarive Hierarchical Clustering (AHC), single lingkage, complete lingkage, average lingkage, cophenetic distance* dan hal-hal lain yang mendukung pendalaman materi.
3. Melakukan pencarian data yang akan dikelompokkan.
4. Merancang aplikasi untuk melakukan pengelompokan data dan mengimplementasikannya ke dalam perangkat lunak.
5. Melakukan pengujian sistem dengan menggunakan data yang diperoleh.
6. Melakukan analisis hasil pengelompokan data.
7. Menentukan kesimpulan dari hasil implementasi dan analisis.
8. Penyusunan laporan Tugas Akhir.

1.5 Sistematika Penulisan

Penulisan Tugas Akhir ini dibagi dalam lima bab, yang terdiri atas :

- Bab 1 Pendahuluan
  Menjelaskan mengenai latar belakang dari pembuatan Tugas Akhir ini, rumusan masalah yang akan dianalisa, batasan dari masalah yang timbul, tujuan yang ingin dicapai dan penentuan metodologi penyelesaian masalah dari sistem yang akan dibuat serta sistematika pembahasan.
- Bab 2 Landasan Teori
  Mengemukakan berbagai teori dasar yang mendukung Tugas Akhir ini, antara lain mengenai *data mining, clustering, agglomerative hierarchical clustering*, dan *cophenetic distance*.
- Bab 3 Analisa dan Perancangan Sistem
  Membahas tentang analisis dan perancangan awal sistem yang akan dibangun dengan tujuan untuk memahami secara jelas proses yang dilakukan pada sistem dalam bentuk Diagram Aliran Data (DAD).
- Bab 4 Implementasi dan Pengujian
  Membahas kebutuhan perangkat lunak dan perangkat keras yang digunakan untuk merealisasikan sistem, membahas scenario pengujian perangkat lunak, hasil uji coba dan analisa dari hasil yang diperoleh.
- Bab 5 Kesimpulan dan Saran
  Berisi kesimpulan dan saran terhadap pengembangan dari penelitian Tugas Akhir ini selanjutnya.

BAB 5
KESIMPULAN DAN SARAN

5.1. Kesimpulan

Kesimpulan yang dapat diambil dari Tugas Akhir ini adalah :

1) Metode *Agglomerative Hierarchical Clustering* (AHC) dengan pendekatan jarak *single linkage, complete* linkage dan *average* linkage dapat digunakan untuk membangun hirarki dari data dan mengelompokkannya.
2) Performansi metode *Agglomerative Hierarchical Clustering* (AHC) dengan pendekatan *average link* pada dataset Iris lebih baik bila dibandingkan dengan *Agglomerative Hierarchical Clustering* (AHC) dengan pendekatan *single link* dan *complete link* yaitu 90,66% berbanding 68% untuk *single linkage* dan 84% untuk *complete link.*
3) Metode pendekatan jarak (*proximity*) sangat berpengaruh dalam membangun hirarki klaster karena perbedaan metode ini menyebabkan hirarki yang dibangunnya pun berbeda.
4) Berdasarkan nilai CPCC yang diperoleh, hasil hirarki metode *Agglomerative Hierarchical Clustering* (AHC) dengan pendekatan *average linkage* lebih baik dibandingkan dengan *Agglomerative Hierarchical Clustering* (AHC) dengan pendekatan *single linkage* dan *complete linkage*.

5.2. Saran

Saran terhadap pengembangan yang akan dilakukan terhadap TA ini adalah :

1) Menggunakan metode clustering lain untuk melakukan pengelompokan data.
2) Menggunakan tipe data lain dalam mengimplementasikan metode *Agglomerative Hierarchical Clustering* (AHC) ini.

Referensi

[1] Borgatti, Stephen P. How To Explain Hierarchical Clustering. Artikel.University of South Carolina. 1994. http://www.analytictech.com/networks/hiclus.htm [12 Maret 2008]

[2] Han, Jiawei, Micheline Kamber. Data Mining: Concepts and Techiniques. Morgan Kaufmann Publishers.2000.

[3] http://en.wikipedia.org/wiki/Data_clustering [12 Maret 2008]

[4] http://lecturer.eepis-its.edu/~tessy/lecturenotes/datamining/chapter10.pdf [20 Agustus 2008]

[5] http://www2.cs.uregina.ca/~dbd/cs831/notes/clustering/clustering.html [20 Maret 2008]

[6] Pramudiono, Iko. Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data. IlmuKomputer.Com. 2003. [10 maret 2008]

[7] Salvador, Stan dan Philip Chan. Determining the Number of Clusters/segments in Hierarchical Clustering/Segmentation Algorithms. Department of Computer Science,Florida Institute of Technology, Melbourne.

[8] Sander,Jorg, Xuejie Qin, Nan Niu dan Alex Kovarsky. Automatic Extraction of Clusters from Hierarchical Clustering Representations. Department of Computing Science,Univercity of Alberta, Canada.

[9] Szymkowiak, A., Larsen, J. and Hansen, L. K. Hierarchical clustering for data mining. Technical University of Denmark, Denmark. 2001.

[10] Tan, Michael. Cluster Analysis of Stock Return. Apothem Capital Management. New York.2002. http://www.michaeltanphd.com/ClusterAnalysisOfStockReturns.pdf [12 Maret 2008]

[11] Vipin Kumar dan Tan Pang Nim. Introduction to Data Mining. Pearson Addison Wesley.

[12] _____. A Tutorial on Clustering Algorithms: Hierarchical clustering algorithm.Artikel. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html [12 Maret 2008]

[13] _____. Agglomerative Hierarchical Clustering Methode. Slide. http://www.bus.utk.edu/stat/Stat579/Hierarchical_Clustering_20Methods.pdf [20 Agustus 2008]