ANALISIS IMPLEMENTASI ALGORITMA MIN-MIN ROUGHNESS (MMR) BERBASIS ROUGH SET THEORY DALAM CLUSTERING PADA DATA KATEGORIKAL

Tegar Arif[1], Agus Nursikuwagus S.t.[2], M.m.[3]

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

Abstrak

Clustering adalah suatu teknik dalam data mining yang bertujuan untuk membagi dan mengelompokkan suatu data berdasarkan kemiripan dari karakteristik pada data tersebut. Telah banyak algoritma untuk menerapkan clustering ini pada data numerikal. Namun pada data kategorikal, yaitu data yang bersifat kategoris dimana suatu objek memiliki value atau isi pada tiap atribut dimana tiap atribut memiliki himpunan value, masih belum banyak. Salah satu algoritma yang dapat menangani clustering pada data kategorikal adalah Min Min Roughness (MMR). MMR memanfaatkan teori rough set dan menggunakan satu parameter input berupa jumlah cluster yang ingin dibentuk, dimana teori rough set ini akan mendapatkan nilai roughness dari tiap value yang ada pada tiap atribut terhadap atribut lain. Dari nilai roughness pada tiap value pada suatu atribut, akan didapatkan nilai mean roughness nya yang selanjutnya akan digunakan dalam pemilihan atribut dan value dari atribut tersebut sebagai dasar dalam pemecahan data. Dari cluster-cluster yang terbentuk akan dianalisis tiap cluster tersebut dengan menggunakan metode purity dan berdasarkan objek-objek yang terdapat dalam cluster tersebut.

Kata Kunci : Clustering, MMR, Teori Rough Set

Abstract

Clustering is one of data mining techniques that splits and classifies data based on similarity from the characteristic of the data. There have been many algorithms that can apply clustering on numerical data. But on categorical data, which is data that categorical where an object holds certain value on specific attribute and each attributes owns set of value, is still few. One of algorithms that can apply clustering on categorical data is Min Min Roughness (MMR). MMR makes use of rough set theory with the number of clusters to be formed as an input parameter, where rough set theory will obtain roughness of each value on certain attribute toward the other ones. From each value roughness on certain attribute, mean roughness will be obtained that will be used to choose attribute and its value as a basis on the splitting of data. From the clusters, will be analyzed on each clusters with purity method and the objects that belong on the cluster.

Keywords : Clustering, MMR, Rough Set Theory

1. Pendahuluan

1.1. Latar Belakang

Clustering adalah salah satu teknik yang terdapat dalam data mining. *Clustering* adalah suatu proses pengelompokkan obyek baik fisik maupun abstrak ke dalam suatu kelas atau *cluster* yang berisi kumpulan obyek yang *similarity* nya tinggi. [2]

Permasalahan pada metode *clustering* yang telah ada adalah kebanyakan hanya dapat diterapkan pada data yang bersifat numerik, karena untuk *clustering* pada data numerik relatif lebih mudah dalam penetapan *similarity* dari *geometric position* nya [1]. Sedangkan untuk data yang bersifat kategorikal, yaitu data yang mempunyai nilai suatu himpunan kategori [3], misal pada atribut *color*, *value* nya adalah *blue*, *white*, *black*, akan sulit diukur nilai *similarity* nya, sehingga dapat terjadi kesalahan dalam memasukkan data tersebut dalam suatu *cluster*. Maka dari itu, pada data kategorikal, akan digunakan metode *clustering* yang berbeda dengan data yang bersifat numerik.

Terdapat beberapa algoritma *clustering* untuk data kategorikal yang telah dirancang sebelumnya seperti CACTUS dan ROCK, namun pada penerapannya di dunia nyata, algoritma-algoritma tersebut masih belum dapat menangani *uncertainty* (ketidakpastian) pada proses clustering. Pada perkembangannya, telah di desain metode *clustering* yang menangani *uncertainty* ini dengan menerapkan himpunan *fuzzy* pada proses *clustering* nya seperti pada fuzzy k-modes dan fuzzy centroids, namun algoritma tersebut membutuhkan beberapa kali pengujian agar menghasilkan nilai parameter yang tepat sehingga dapat mencapai kondisi stabilitas yang ingin dicapai.

Terdapat suatu algoritma yang dirancang untuk dapat memecahkan masalah diatas, yaitu Min-Min Roughness (MMR). Algoritma MMR ini dirancang agar dapat menangani *uncertainty* pada saat proses *clustering* data kategorikal dengan menerapkan penggunaan Rough Set Theory (RST). RST adalah suatu metode untuk mendapatkan *decision making* pada kemunculan *uncertainty* dengan memakai konsep *lower approximation* dan *upper approximation* yang selanjutnya digunakan untuk mendapatkan nilai *roughness* pada tiap atribut terhadap atribut-atribut lainnya. Selanjutnya, algoritma MMR mencari nilai *roughness* yang minimal pada tiap-tiap atribut tersebut, dan memproses ke seluruh atribut yang ada sehingga didapatkan nilai *roughness* minimal untuk tiap-tiap atribut berdasarkan seluruh atribut lainnya. Langkah selanjutnya adalah melakukan pembagian (*split*) terhadap kumpulan objek sampai didapatkan jumlah *cluster* yang sesuai dengan parameter.

Pada tugas akhir ini akan mengimplementasikan algoritma Min-Min Roughness (MMR) untuk melakukan *clustering* pada data kategorikal yang berbasis Rough Set Theory (RST) dan menganalisis nilai akurasi dari hasil yang didapatkan.

1.2. Perumusan Masalah

Beberapa masalah yang dapat diangkat dari tugas akhir ini adalah :

1. Bagaimana penerapan algoritma MMR ini dalam menyelesaikan permasalahan *clustering* data kategorikal ?
2. Bagaimana paramater jumlah cluster mempengaruhi nilai *purity* dari hasil *clustering* ?

Batasan masalah dari tugas akhir ini adalah :

1. *Dataset* yang digunakan berasal dari UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets.html.
2. *Dataset* bertipe kategorikal dan mempunyai informasi di kelas mana tiap objek berada.

1.3. Tujuan

Tujuan dari tugas akhir ini adalah

1. Mengimplementasikan algoritma MMR dalam proses *clustering* data kategorikal.
2. Mengevaluasi nilai *purity* dari hasil proses *clustering* berdasarkan jumlah cluster yang terbentuk.

1.4. Metodologi Penyelesaian Masalah

Metodologi yang digunakan pada tugas akhir ini adalah :

1. Studi Literatur
   Pencarian sumber-sumber dan referensi literatur yang berkaitan dengan data mining, *clustering*, data kategorikal, algoritma MMR dan Rough Set Theory.
2. Pengumpulan Data
   Mempersiapkan *dataset* berupa data bertipe kategorikal yang akan digunakan dalam proses *clustering.*
3. Perancangan dan Implementasi Sistem
   Perancangan sistem yang akan diimplementasikan yaitu penggunaan algoritma MMR dalam proses *clustering* pada data kategorikal berbasis Rough Set Theory, penentuan bahasa pemrograman yang akan digunakan, beserta fungsionalitas serta antarmuka.
4. Pengujian dan Analisis Hasil
   Pengujian dari sistem yang telah diimplementasikan dengan data yang ada selanjutnya dilakukan analisis terhadap jumlah *cluster* yang terbentuk dan nilai *purity* dari hasil proses *clustering* tersebut.
5. Penyusunan Laporan Tugas Akhir
   Penyusunan laporan dan dokumentasi dalam pembuatan tugas akhir sesuai dengan kaidah dan tata cara yang telah ditetapkan.

5. Kesimpulan Dan Saran

5.1. Kesimpulan

Berdasarkan pengujian yang telah dilakukan pada bab sebelumnya, dapat disimpulkan bahwa algoritma MMR dapat melakukan *cluster*ing pada data kategorikal dengan nilai purity diatas 0.7. Namun penambahan jumlah *cluster* lebih dari jumlah kelas asli walaupun menaikkan nilai purity namun *cluster* yang terbentuk rata-rata tidak efektif karena tidak merepresentasikan suatu kelas asli dari dataset.

Kelemahan dari algoritma ini adalah tidak dapat menangani dataset yang jumlah objek tiap kelasnya tidak seimbang (imbalanced class) karena pada saat melakukan pemecahan untuk *cluster* selanjutnya dipilih *cluster* dengan jumlah objek terbanyak, hal ini bisa jadi akan memecah suatu *cluster* dimana *cluster* tersebut telah merepresentasikan suatu kelas asli yang jumlah objeknya memang besar. Algoritma ini juga tidak dapat melakukan *cluster*ing pada data yang *value* nya bersifat kombinatorial seperti pada dataset Car, karena nilai mean roughness dari tiap atribut semua bernilai 1 yang berarti tiap atribut tidak memiliki kemiripan apabila menggunakan metode ini. Selain itu algoritma ini juga cukup memakan waktu apabila jumlah datanya besar seperti pada dataset mushroom, karena melakukan perulangan yang cukup banyak untuk membentuk *cluster*.

5.2. Saran

Perlu dilakukan analisis lebih lanjut terhadap hasil kualitas *cluster* yang terbentuk, karena walaupun purity meningkat dengan penambahan jumlah *cluster*, namun *cluster* yang terbentuk kurang efektif. Pemilihan *cluster* mana yang akan dipecah juga tidak berdasarkan *cluster* yang memiliki jumlah objek terbesar, namun perlu analisis lebih lanjut mana *cluster* yang paling tepat untuk dipecah. Dan kedepannya, algoritma MMR ini diharapkan dapat menangani suatu dataset yang bersifat numerikal, tidak hanya kategorikal saja.

Daftar Pustaka

[1] Parmar. Darshit, Wu*.Teresa, Blackhurst.Jennifer,2007,*MMR: An Algorithm For Clustering Categorical Data Using Rough Set Theory,* Department of Industrial EngineeringArizona State University

[2] Han.Jiawei, Kamber.Micheline, 2006,*Data Mining : Concepts and Technique Second Edition,* University of Illionis at Urbana-Champaign

[3] Agresti. Alan, 2007, *An Introduction To Categorical Data Analysis Second Edition*, Department of Statistics University of Florida

[4]Cattaneo, Gianpiero, 2010, *Rough Set Theory II The Approximation Space Approach,*Dipartimento di Informatica, Sistemistica e ComunicazioneUniversitµa di Milano

[5]Hasegawa. Kiyoshi, Koyama. Michio, Arakawa. Masamoto, Funatsu. Kimito, 2009, *Application Of Data Mining To Quantitative Structure-Activity Relationship Using Rough Set Theory,* Chemometrics and Intelligent Laboratory Systems 99 (2009) 66 – 70

[6]Niknam. Taher, Bahmani Firouzi. Bahman, Nayeripour. Majid, 2008, *An Efficient Hybrid Evolutionary Algorithm for Cluster Analysis,* World Applied Sciences Journal 4 (2): 300-307, 2008

[7] Herawan. Tutut, Tri Riyadi Yanto. Iwan, Mat Deris. sMustafa, 2009, *Rough Set Approach For Categorical Data Clustering,* Faculty of Information Technology and Multimedia University Tun Hussein Om Malaysia

[8]J. Mazlack. Lawrence, He. Aijing, Zhu. Yaoyao, Coppock. Sarah, 2000, *A Rough Set Approach in Choosing Partitioning Attributes,*Computer Science University of Cincinnati

[9] Pawlak. Zdzislaw, Grzymala-Busse. Jerzy, Slowinski. Roman, Ziarko. Wojciech, 1995 *Rough Sets,* Communication of The ACM November 1995/Vol. 38, No. 11