ANALISIS DAN IMPLEMENTASI FUZZY CLUSTERING UNTUK FEATURE SELECTION

Okti Purwaningsih[1], Deni Saepudin[2], Intan Nurma Yulita[3]

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

Abstrak

Data yang akan diolah untuk keperluan mencari informasi tertentu terkadang tidak semuanya relevan. Terdapat beberapa feature yang irrelevan maupun redundan. Feature-feature tersebut perlu ditangani. Salah satu cara menanganinya adalah dengan cara feature selection. Feature selection yaitu proses mengurangi dimensi dengan menghapus feature-feature yang tidak relevan. Pada tugas akhir ini dilakukan feature selection menggunakan fuzzy clustering yaitu fuzzy c-means dengan pendekatan wrapper. Tujuan dari feature selection ini adalah struktur dari subset feature memiliki kemiripan yang hampir sama dengan struktur dari data asli yang dilihat dari cluster-cluster yang terbentuk oleh subset feature dan data asli. Kemiripan tersebut dapat dilihat dengan classification error rate, dimana semakin kecil nilainya maka semakin mirip struktur subset feature dengan data asli. Metode ini mampu memilih subset feature yang memiliki kesamaan struktur dengan dataset asli. Bahkan classification error rate dataset subset feature tidak meningkat bahkan berkurang jika subset feature yang terpilih adalah relevan.

Kata Kunci : feature selection, Fuzzy C-Means, wrapper feature selection, subset feature

Abstract

In order to search some certain information, we don't always get all relevant data, there are some irrelevant and redundant features. This matter should be fix properly by using feature selection. Feature selection is a reducing process of dimension by deleting some irrelevant features. This final project will provide a research explanation of making feature selection using fuzzy clustering, that is fuzzy c-means, by using wrapper approach. The structure of the feature subset is identical to the real data structure from the cluster that made by feature subset and real data. This identical structure can be known from classification error rate, the smaller the value, more identical the feature from the real data. This method can selectively choose the feature subset that have identical structure with the real data set. Even the classification error rate of data set feature subset will decrease if the chosen feature subset is relevant.

Keywords : feature selection, Fuzzy C-Means, wrapper feature selection, subset feature

BAB 1 Pendahuluan

1.1 Latar Belakang

Seiring perkembangan zaman, jumlah data semakin berkembang. Namun, pada data-data tersebut mengandung feature yang tidak relevan serta feature yang redudansi yang menyebabkan dimensinya bertambah. Padahal tidak semua feature-feature tersebut relevan untuk diolah. Oleh sebab itu, perlu adanya *preprocessing* data untuk mengurangi jumlah feature. Salah satu caranya adalah *feature selection*. *Feature selection* yaitu proses mengurangi dimensi dengan menghapus feature-feature yang tidak relevan [3]. Hal ini perlu dilakukan untuk menghapus featur-feature yang tidak relevan atau berlebihan sehingga akan meningkatkan akurasi pada klasifikasi. Selain itu, dengan jumlah feature yang lebih sedikit akan mempercepat proses pembentukan model.

Penelitian mengenai *feature selection* telah banyak dilakukan selama beberapa tahun. Berbagai macam metode *feature selection* telah dirangkum dan dijabarkan [2]. Beberapa pendekatan yang sering digunakan untuk *feature selection* diantaranya pendekatan *filter* dan *wrapper*. Strategi pencarian *subset feature*-nya pun bermacam-macam seperti *complete*, heuristik, dan random. Terdapat 5 tipe fungsi untuk mengevaluasi subset feature, diantaranya adalah *distance measure*, *information measure*, *dependence measure*, *consistency measure*, dan *classification error rate measure* [2].

Pendekatan *feature selection* yang digunakan pada penelitian ini berdasarkan struktur pada subset feature yang memiliki kemiripan struktur atau setidaknya mendekati struktur pada original dataset. Kemiripan struktur subset feature dan original dataset akan dideskripsikan menggunakan *cluster* yang dapat dilihat tingkat kemiripannya dengan *classification error rate*. Oleh karena itu, pada tugas akhir ini akan dilakukan proses *feature selection* menggunakan *fuzzy clustering* untuk klasifikasi dengan metode *wrapper*. Metode *wrapper* memiliki tingkat keakuratan yang lebih tinggi dibandingkan menggunakan metode *filter* karena menggunakan akurasi sebenarnya dalam mengevaluasi subset feature. Evaluasi subset feature menggunakan *classification error rate* dengan induksi algoritma *fuzzy clustering* (*fuzzy c-means*). *Fuzzy clustering* merupakan salah satu metode yang telah berhasil mengurangi jumlah feature, mengurangi waktu komputasi dan memori, meningkatkan akurasi dengan rata-rata *classification error rate*-nya menurun hingga 23 % dari 44,6 % atau mengalami penurunan *error rate* sebanyak 21,6 % [10,11]. *Fuzzy clustering* merupakan metode *prototype-based clustering* yang mengadopsi teori *fuzzy set* dimana sebuah objek dapat dimiliki oleh beberapa *cluster* dengan memiliki *membership degree* pada tiap *cluster*. Hal ini merupakan kelebihan dari *fuzzy clustering* dimana suatu objek yang berada diantara dua *cluster* atau lebih sulit untuk ditentukan *cluster*-nya jika menggunakan *hard clustering*.

1.2 Perumusan Masalah

Berdasarkan pada latar belakang di atas, permasalahan yang akan diuraikan dan diteliti adalah :

1. Merancang penentuan suatu feature untuk terpilih atau tidak.

2. Mengimplementasikan fuzzy clustering untuk menghitung derajat keanggotaan feature pada setiap *cluster*.
3. Membandingkan performansi pada data yang telah dilakukan proses *feature selection* dan data yang yang belum dilakukan *feature selection.*

1.3 Batasan Masalah

Dalam implementasi Tugas Akhir ini akan dibatasi oleh beberapa hal, yaitu :

1. Dataset yang digunakan merupakan dataset yang dipublikasikan melalui internet.
2. Dataset yang digunakan merupakan dataset untuk kasus *single label* dan bertipe numerik.

1.4 Tujuan

Tujuan yang ingin dicapai dalam penelitian ini adalah :

1. Mengimplementasikan *feature selection* dengan *fuzzy clustering* untuk mengurangi jumlah feature pada suatu data untuk mendapatkan feature yang relevan.
2. Menganalisis performansi data setelah dilakukan *feature selection* dan sebelum dilakukan *feature selection* dengan *classification error rate* dan waktu pemrosesan.

1.5 Metode Penyelesaian Masalah

Metodologi yang digunakan dalam memecahkan permasalahan-permasalahan dalam Tugas Akhir ini terdiri dari beberapa tahap berikut :

1. Studi literatur.

Pada tahap ini dilakukan pencarian referensi-referensi untuk penelitian Tugas Akhir seperti referensi tentang konsep *feature selection, fuzzy clustering, clustering*, dan materi pendukung lainnya.

2. Pengumpulan Data

Pada tahap ini dilakukan pencarian data-data pendukung seperti dataset yang digunakan untuk melakukan penelitian pada Tugas Akhir ini. Dataset ini dicari pada website penyedia layanan dataset untuk *machine learning* dengan memperhatikan bahwa *single label* dan bertipe numerik.

3. Analisis dan perancangan kebutuhan sistem.

Melakukan analisis dan perancangan sistem yang akan dibangun. Tentunya sistem yang akan dibangun sesuai dengan metode yang digunakan pada Tugas Akhir ini yaitu *fuzzy clustering*.

4. Implementasi Sistem

Pada tahap ini dilakukan pembangunan sistem menggunakan tool Matlab R2008a sesuai dengan rancangan pada tahap sebelumnya.

5. Pengujian sistem dan Analisis hasil pengujian.

Pada tahap ini akan dilakukan pengujian sistem apakah program yang dibuat tersebut sudah sesuai dengan fungsionalitas yang telah dirancang pada tahap perancangan sistem. Selain itu, pada tahap ini juga dilakukan analisis hasil *feature selection* untuk kasus klasifikasi dari segi akurasi maupun waktu eksekusi dari metode *fuzzy clustering*.

6. Penyusunan laporan Tugas Akhir.

Pada tahap ini, dilakukan penyusunan laporan akhir dan pengumpulan dokumentasi yang diperlukan, format laporan mengikuti kaidah penulisan yang benar dan yang sesuai dengan ketentuan-ketentuan yang telah ditetapkan oleh institusi.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN

Berisi latar belakang, perumusan masalah, batasan masalah, tujuan, hipotesa, metode penyelesaian masalah dan sistematika penulisan.

BAB II LANDASAN TEORI

Berisi penjelasan singkat mengenai konsep-konsep yang mendukung dikembangkannya sistem ini.

BAB III ANALISIS DAN PERANCANGAN SISTEM

Berisi rincian mengenai desain sistem untuk *feature selection* menggunakan *fuzzy clustering*.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Berisi rincian mengenai pengujian yang dilakukan terhadap sistem yang dikembangkan serta analisis terhadap hasil pengujian.

BAB V KESIMPULAN DAN SARAN

Berisi kesimpulan yang diambil berkaitan dengan sistem yang dikembangkan serta saran-saran untuk pengembangan lebih lanjut.

BAB 5 Penutup

5.1 Kesimpulan

Dari penelitian yang telah dilakukan maka dapat diambil kesimpulan sebagai berikut :

1. *Feature selection* menggunakan *fuzzy c-means* dengan metode *wrapper* dapat mengurangi jumlah feature, karena memilih *subset feature* yang memiliki classification error terbaik secara *sequential forward selection*. Namun, pengurangan jumlah feature ini berdampak pada kualitas dataset dengan *subset feature* karena penggunaan *wrapper method* yang rentan terhadap overfitting. Jika subset feature yang terpilih adalah feature-feature relevan, maka *classification error rate*nya akan baik, tetapi jika tidak relevan bahkan terlalu sedikit maka *classification error rate*nya akan menjadi buruk.
2. Metode ini mampu memilih *subset feature* dimana memiliki kesamaan struktur dengan dataset asli dengan membandingkan hasil *classification error rate*. Jika *classification error rate* antara dataset *subset feature* dan dataset original sama, maka struktur data kedua dataset adalah sama.
3. Nilai bobot pada *fuzzy c-means* berpengaruh pada pemilihan subset feature. Jika karakteristik datasetnya menyebar, maka semakin besar nilai bobot akan menghasilkan *classification error rate* yang lebih baik. Sedangkan jika karakteristik datasetnya mengumpul, maka semakin kecil nilai bobot akan menghasilkan *classification error rate* yang lebih baik.
4. *Classification error rate* dataset *subset feature* tidak meningkat bahkan berkurang jika *subset feature* yang terpilih adalah relevan. Pemilihan *subset feature* kurang baik jika diimplementasikan pada dataset yang target kelasnya dari feature-feature yang berhubungan atau suatu feature tidak dapat dipisah-pisah proses evaluasinya secara individu karena pemilihan feature dengan metode ini memilih feature yang memiliki *class correlation* tertinggi meskipun belum tentu feature ini yang relevan.
5. *Running time feature selection* tidak lebih baik daripada tanpa *feature selection* karena penggunaan pendekatan *wrapper* yang menghabiskan biaya lebih banyak dan dipengaruhi oleh jumlah subset feature yang dievaluasi.

5.2 Saran

Saran yang diberikan penulis berkaitan dengan penelitian yang telah dilakukan adalah :

1. Pada penelitian ini menggunakan metode *searching sequential forward selection* yang bersifat *greedy*, untuk menghindari terjadinya *local minimum* dapat dicoba dengan metode *best first search* yang memiliki kemampuan *backtracking*.

Daftar Pustaka

[1] Bezdek, James C, Robert Ehrlich, William Full. 1984. "FCM : The Fuzzy C-Means Clustering Algorithm". Computers and Geosciences Vol. 10, No. 2-3, pp. 191-203, 1984

[2] Dash, M. dan H. Liu. 1997. "Feature Selection for Classification". Intelligent Data Analysis 1 (1997) 131–156

[3] Guyon, I., dan Elisseeff, A. 2003. "An Introduction to Variable and Feature Selection". JMLR 5 : 845 – 889

[4] Han, Jiawei. dan Micheline Kamber. 2006. *Data Mining Concepts and Techniques 2nd Edition.* Morgan Kaufmann.

[5] Hoppner, Frank, Frank Klawonn, Rudolf Kruse, Thomas Runkler. 1999. *Fuzzy Cluster Analysis : Methods for Classification, Data Analysis and Image Recognition.* England: John Wiley & Sons Ltd

[6] Jihong, Pei, Xuan, Yang, Xinbo, Gao, Weixin, Xie. 2001. "On The Weighting Exponent m in Fuzzy C-Means (FCM) Clustering Algorithm". Proceeding of SPIE Vol. 4554 (2001)

[7] Kohavi, Ron, dan John, George H., 1997. "Wrappers for Feature Subset Selection". Elsevier Science Artificial Intelligence 97 (1997) 273 – 324

[8] Kohavi, Ron, dan John, George H., 1997. "The Wrapper Approach".

[9] Kusumadewi, Sri; Purnomo, Hari. 2004. *Aplikasi Logika Fuzzy untuk Pendukung Keputusan*. Penerbit Graha Ilmu

[10] Sun, Hao Jun, Mei Sun, Zhen Mei. 2006. "Feature Selection via Fuzzy Clustering". Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006

[11] Sun, Hao Jun,Sheng-Rui Wang, Zhen Mei. 2002. "A Fuzzy Clustering Based Algorithm for Feature Selection". Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, 4-5 November 2002

[12] Tan, Pang-Nin, Steinbach, Michael, Kumar, Vipin. 2006. *Introduction to Data Mining.* Boston : Pearson Education Inc.