IMPLEMENTASI DAN ANALISIS FEATURE SELECTION MENGGUNAKAN MUTUAL INFORMATION AND REDUNDANCY-SYNERGY COEFFICIENT
IMPLEMENTATION AND ANALYSIS OF FEATURE SELECTION USING MUTUAL INFORMATION AND REDUNDANCY-SYNERGY COEFFICIENT

Putri Karima Hapsari[1], Imelda Atastina[2], Kusuma Ayu Laksitowening[3]

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

Abstrak

Preprocessing data adalah proses pembersihan data dari noise, outlier, missing value, irrelevant feature, redundant feature dll agar data yang digunakan bisa lebih efisien saat dilakukan proses data mining selanjutnya. Salah satu cara preprocessing adalah dengan mereduksi dimensi data. Metode reduksi data yang bisa dilakukan adalah dengan memilih feature-feature yang penting saja (feature selection). Pada tugas akhir ini diterapkan metode feature selection MMIMRSC (Maintaining mutual information and minimizing redundancy-synergy coefficient) yang bertujuan menghilangkan feature irelevan dan redundan dengan menghitung nilai informasi feature serta mengurangi redundansi feature dengan menghitung redundancy-synergy coefficientnya. Hasil performansi data hasil preprocessing, yaitu nilai precision dan recall nya mengalami peningkatan dan waktu pembentukan model klasifikasi menjadi lebih cepat.

Kata Kunci : feature selection, mutual information, reduksi data, preprocessing

Abstract

Preprocessing data is a process to cleaning data from noise, outlier, missing value, irrelevant feature, redundant feature etc, so that the data can be more efficient when the next mining process implemented. One of preprocessing techniques is by reduction the dimension of data. Data reduction method that can be implemented is feature selection (choosing only the important features). In this TA, feature selection method that being implemented is MMIMRSC (Maintaining mutual information and minimizing redundancy-synergy coefficient) that purpose to discard irrelevant and redundant feature by calculating the information value and also decreasing the redundant feature by calculating redundancy-synergy coefficient. The performance of reduces data after preprocessing shows that precision and recall value increase and time to build classification model become faster than unreduced data.

Keywords : feature selection, mutual information, data reduction, preprocessing

BAB I

PENDAHULUAN

1.1 Latar belakang masalah

Pada era informasi sekarang ini, media penyimpanan dan teknologi retrieval di internet semakin banyak digunakan, sehingga data yang disimpan bertambah dari jumlah maupun dimensi [1]. Pada umumnya, data yang dimiliki sering kali mengandung *noise, missing value, redundant* dan *irrelevant* feature yang bisa membuat pengolahan data kurang efisien. Salah satu solusi yang bisa ditawarkan untuk menghasilkan data yang berkualitas adalah dengan memilih *feature-feature* penting saja untuk diproses.

Feature subset selection (FSS) merupakan masalah umum yang terdapat pada *data mining* mengenai bagaimana memilih features yang relevan dan menghilangkan features yang redundant (berulang) dari feature set yang asli [2]. Semakin kecil ukuran feature subset, makin optimal feature subset yang didapat. Jika feature dipilih secara tepat maka dapat meningkatkan akurasi dan performansi dari proses klasifikasi [3]. Karena itu, *feature selection* menjadi sangat perlu dilakukan untuk mendukung *preprocessing* untuk *machine learning*.

Feature selection untuk preprosess data sudah cukup banyak diterapkan. Feature selection bisa dibagi dalam tiga jenis, yaitu *filter model*, *wrapper model* dan *embedded model*. Contoh-contoh penerapan model filter antara lain Mutual Information [6] dan Relief algoritma [5]. Mutual information (MI) yang menggunakan pendekatan information meassure sangat efektif untuk mengevaluasi relevansi dari setiap variabel input namun gagal untuk mengeliminasi feature yang redundan, begitu pula pada relief algoritma yang menggunakan pembobotan untuk melihat feature mana yang tidak relevan namun tidak mempertimbangkan redundancy data.

Salah satu teknik *feature subset selection* yang dapat digunakan adalah *Mutual Information and Redundancy-Synergy Coefficient*. Metode ini merupakan salah satu model berpendekatan *information meassure* yang merupakan pengembangan dari metode mutual infomation, namun sudah dilengkapi dengan cara mengurangi feature yang redundan. Pada penerapan mutual information sebelumnya [6], memungkinkan untuk menghilangkan feature yang tidak relevan tapi tidak bisa mengurangi feature yang redundan.

Terkadang feature redundan dihilangkan namun menghilangkan feature yang sebenarnya relevan [3]. Pada metode ini akan dihitung feature subset mana yang memiliki nilai informasi yang tinggi. Setelah itu diterapkan Redundancy-Synergy coefficient untuk mengukur tingkat *redundancy* data tanpa mengurangi tingkat relevansi data. Jika pada paper sebelumnya[2] yang diamati lebih menekankan pada kecepatan proses *preprocessing* dan tingkat akurasi umum, maka pada tugas akhir ini lebih diamati hasil tingkat ketepatan hasil klasifikasi masing-masing nilai kelas (diukur dengan precision dan recall) serta waktu pembentukan model saat dilakukan klasifikasi dengan algoritma klasifikasi

Naives Bayes. Algoritma klasifikasi tersebut dipilih karena menggunakan keseluruhan dataset saat pemrosesan dan tidak terdapat preprocess sebelumnya. Selain itu algoritma klasifikasi Naives bayes juga merupakan algoritma klasifikasi yang cepat dalam prosesnya.

Hipotesis awal dari pengerjaan tugas akhir ini adalah, dengan melakukan *preprocessing* menggunakan metode mutual information dan *redundancy-synergy coefficient* dapat meningkatkan nilai *precision* dan *recall* serta waktu pembentukan model saat dilakukan proses klasifikasi.

1.2 Perumusan masalah

Berdasarkan latar belakang yang telah diuraikan di atas, maka dapat dirumuskan beberapa masalah antara lain:

1. Bagaimana melakukan reduksi dimensi data menggunakan *Mutual Information and Redundancy-Synergy coefficient* agar pengolahan data untuk proses selanjutnya lebih baik?
2. Bagaimana performansi reduksi data setelah dilakukan *feature selection* dengan *Mutual Information and Redundancy-Synergy coefficient* dinilai dari parameter uji seperti *precison* dan *recall* serta waktu pembentukan model saat menggunakan algoritma klasifikasi?

1.3 Batasan Masalah

Adapun beberapa batasan masalah dalam pengerjaan TA ini, antara lain:

1. Dataset yang digunakan adalah data dengan tipe atribut diskrit
2. Untuk analisis hasil, parameter yang digunakan untuk mengevaluasi hasil adalah *precision* dan *recall* serta waktu pembentukan model saat hasil *feature selection* diterapkan pada algoritma klasifikasi.
3. Tidak menangani *data cleaning (missing value, noisy data), data integration* dan *data transformation.*
4. Evaluasi dataset hasil reduksi menggunakan algoritma klasifikasi naives bayes

1.4 Tujuan

Tujuan dari pengerjaan tugas akhir ini antara lain adalah:

1. Mereduksi dimensi data menggunakan metode *Mutual Information and Redundancy-Synergy coefficient* sehingga akan lebih efisien saat dilakukan pemrosesan selanjutnya, yaitu proses klasifikasi dengan *naives bayes*
2. Menganalisis performansi hasil preprocessing menggunakan *Mutual Information and Redundancy-Synergy coefficient* menggunakan algoritma klasifikasi dengan parameter-parameter pengukuran *precison, recall* dan waktu pembentukan model saat dilakukan proses klasifikasi.

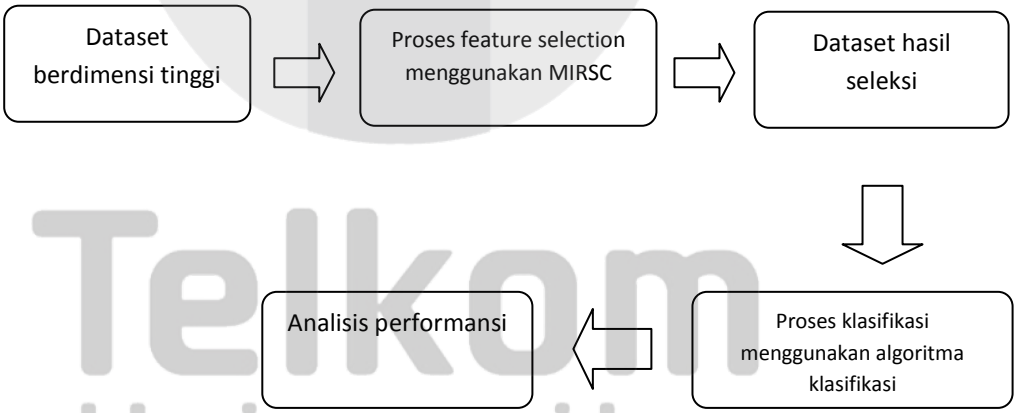
1.5 Metodologi penyelesaian masalah

Dalam pengerjaan tugas akhir ini, digunakan metodologi penyelesaian masalah sebagai berikut:

1. Identifikasi Masalah
   Permasalahan yang muncul dan melatarbelakangi dibuatnya tugas akhir ini antara lain:
   - ❖ Bagaimana memilih feature subset yang minimal agar dataset yang berdimensi besar bisa lebih efisien ketika diproses di task klasifikasi.
   - ❖ Bagaimana menghilangkan feature redundan dan irrelevan pada sebuah dataset berdimensi besar, namun tetap memiliki nilai informasi yang sama dengan dataset asli

2. Study literatur
   Melakukan studi literatur yang berhubungan dengan feature selection, klasifikasi, mutual information dan materi-materi yang berhubungan dengan pengerjaan TA melalui sumber-sumber literatur maupun sumber online.

3. Desain Penelitian
   Pada pengerjaan tugas akhir ini, alur pengerjaan yang akan dilakukan adalah sebagai berikut:



Gambar 1.1 Alur pengerjaan

Keterangan:
1. Dipilih sebuah dataset berdimensi tinggi untuk dilakukan preprocessing

2. Kemudian dataset diproses menggunakan metode mutual information dan redundancy-synergy coefficient agar dapat menghilangkan data yang berulang dan tidak relevan
3. Kemudian dari proses tersebut didapat dataset hasil reduksi
4. Dataset hasil reduksi diproses menggunakan algoritma klasifikasi, yaitu naives bayes
5. Setelah itu, hasil klasifikasi akan dibandingkan dengan hasil pemrosesan dataset sebelum di preprosess agar dapat dilihat perbandingan *precision*, *recall* serta waktu pembentukan model saat dilakukan klasifikasi.

Aplikasi diimplementasikan menggunakan Matlab, proses klasifikasi akan diuji menggunakan Weka sebagai alat bantu. Algoritma klasifikasi yang digunakan untuk pengujian adalah Naives Bayes

4. Analisis
   Hipotesis awal dari pengerjaan Tugas akhir ini adalah, dengan melakukan *preprocessing* menggunakan metode mutual information dan redundancy-synergy coefficient dapat meningkatkan *precison*, *recall* dan waktu pembentukan model saat dilakukan proses klasifikasi.

1.5 Sistematika Penulisan

Penulisan tugas akhir ini dibagi ke dalam 5 bab, antara lain :

1. **Bab 1 Pendahuluan**
   Berisi latar belakang masalah, perumusan masalah, batasan masalah, tujuan masalah, metodologi penyelesaian masalah dan sistematika penulisan

2. **Bab 2 Landasan Teori**
   Berisi landasan teori tentang KDD, metode *mutual information* dan *redundancy-synergy coefficient* dan klasifikasi.

3. **Bab 3 Perancangan Aplikasi**
   Berisi analisa dan perancangan aplikasi yang akan dibuat dengan bahasa permodelan DFD

4. **Bab 4 Analisa Hasil**
   Berisi penjelasan mengenai implementasi hasil perancangan, uji coba terhadap sistem, dan analisa perangkat lunak yang dibangun.

5. **Bab 5 Kesimpulan dan Saran**
   Berisi kesimpulan dan saran dari hasil penelitian untuk pengembangan lebih lanjut.

BAB V

SIMPULAN DAN SARAN

5.1 Simpulan

Dari hasil penelitian di atas dapat diambil beberapa kesimpulan sebagai berikut:

1. Hasil *preprocessing* dipengaruhi oleh karakteristik dataset seperti variasi nilai feature (tingkat homogenitas variasi nilai feature) dan nilai mutual informasi feature
2. Waktu pembentukan model setelah preproses menjadi lebih cepat karena hanya feature yang memiliki nilai informasi tinggi yang diikutkan. Namun karena waktu pembentukan model klasifikasi Naives bayes sudah cepat (hampir mendekati 0 pada dataset uji), maka setelah preproses tidak dapat diamati peningkatan waktu yang signifikan karena perbedaan antara sebelum dan sesudah preproses sangat kecil.
3. Untuk dataset dengan feature kecil dan instances sedikit, proses MMIMRSC cukup efisien karena waktu preprocess kecil, sedangkan untuk dataset dengan feature banyak dan instances banyak, proses kurang efisien dari segi waktu karena pemrosesan relative lama. Namun *preprocessing* menggunakan MMIMRSC dapat meningkatkan akurasi dari proses klasifikasi.

5.2 Saran

1. Dilakukan pengujian juga terhadap data yang kontinu
2. Data yang memiliki *missing value* harusnya ditangani di sistem MMIMRSC
3. Diamati batas homogenitas variasi nilai yang memungkinkan akurasi dataset meningkat atau menurun

DAFTAR PUSTAKA

[1] Han, Jiawei. Kamber, Micheline. 2001. *Data Mining Concepts and Techniques*. San Francisco: Morgan Kaufmann Publisher. Inc

[2] Sheng, Yang, Jun, Gu. 2004. *Feature selection based on mutual information and redundancy-synergy coefficient,* [online], (www.zju.edu.cn/jzus/2004/0411/041111.pdf, diakses tanggal 11 Febuari 2010)

[3] Bonev, Boyan, Escolano, Francisco, and Miguel Angel Cazorla. 2007. *A Novel Information Theory Method for Filter Feature Selection*, [online], (http://www.dccia.ua.es/~boyan/papers/micai07bonev_draft.pdf, diakses tanggal 11 febuari 2010)

[4] Pressman, Roger.2005. *Software Engineering: A Practitioner's Approach sixth edition*. New York: McGraw-Hill Companies, Inc.

[5] Yu, Ley, Huan Liu. 2003. *Feature Selection doh High-Dimensional Data: A Fast Correlation Based Filter Solution,* [online], (http://www.hpl.hp.com/conferences/icml2003/papers/144.pdf, diakses tanggal 02 Febuari 2010)

[6] Hua Yang, Howard, Moody, John. 1999. *Feature Selection Based on Joint Mutual Information*, [online], (https://eprints.kfupm.edu.sa/40974/1/40974.pdf, diakses tanggal 11 Febuari 2010)

[7] Sebban, Marc. 1999. *On Feature Selection: a New Filter Model*, [online], (http://www.aaai.org/Papers/FLAIRS/1999/FLAIRS99-041.pdf, diakses tanggal 11 Febuari 2010)

[8] El Akadi, Ali, El Ouardighi, Abdeljalil and Aboutajdine, Driss. 2008. *A Powerful Feature Selection approach based on Mutual Information*, [online], (http://paper.ijcsns.org/07_book/200804/20080417.pdf, diakses tanggal 11 Febuari 2010)

[9] Santoso, Budi.2007. *Data Mining: Teknik Pemanfaatan Data Untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.

[10] Kusrini dan Lutfhi, Emha Taufiq. 2009. *Algoritma Data Mining*. Yogyakarta: Penerbit ANDI.