*Evaluasi Seleksi Fitur Berbasis Filter dan Wrapper pada Diagnosa Penyakit Kanker Payudara*

Nama

Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang, Indonesia

email@gmail.com

*Abstrak*—Penelitian ini bertujuan untuk melakukan evaluasi penggunaan seleksi fitur berbasis filter dan wrapper untuk diagnosa penyakit kanker payudara. Algoritma Naïve Bayes dipilih sebagai algoritma klasifikasi karena memiliki kompleksitas algoritma yang rendah. Penelitian ini menunjukan bahwa algoritma seleksi fitur tersebut mampu meningkatkan algoritma Naïve Bayes dalam melakukan diagnose penyakit kanker payudara. Tingkat akurasi yang paling baik ditunjukan oleh penggunaan BFE dan CHI sebagai algoritma seleksi fitur. Walaupun secara keseluruhan, penggunaan seleksi fitur berbasis filter dan wrapper dalam algoritma Naïve Bayes masih memiliki akurasi yang lebih rendah dibandingkan dengan penelitian sebelumnya yang menggunakan dataset dari UCI Machine Learning dataset.

Kata Kunci—chi square, forward feature selection, backward feature elimination, naïve bayes

#  Pendahuluan

Penelitian ini membahas tentang diagnosa penyakit kanker payudara. Dari seluruh penderita kanker di Mesir, hampir 18,3% menderita penyakit kanker payudara [[1](#Sal12)]. Dengan kondisi tersebut, penelitian ini memandang perlu untuk melakukan penelitian mengenai diagnosa penyakit payudara.

Dalam melakukan penelitian ini, penulis merujuk pada dua paper sebelumnya yang keduanya menggunakan dataset yang sama. Chunekar dan Ambulgekar [[2](#Chu09)] melakukan evaluasi terhadap penggunaan algoritma Neural Network untuk diagnosa penyakit kanker payudara. Dengan menggunakan dataset WDBC yang berasal dari UCI, penelitian tersebut menghasilkan akurasi sebesar 70,725%.

Penelitian yang kedua dilakukan oleh Salama, dkk [[1](#Sal12)] melakukan uji perbandingan terhadap algoritma decision tree (J48), Multi-Layer Perception (MLP), Naive Bayes (NB), Sequential Minimal Optimization (SMO), dan Instance Based untuk K-Nearest neighbor (IBK) pada diagnosa penyakit kanker payudara. Dalam perbandingan tersebut, SMO menghasilkan akurasi yang paling baik, yaitu sebesar 97,71%. Penelitian tersebut juga melakukan penggabungan metode antara SMO+J48-MLP+IBk. Namun, untuk dataset WDBC, akurasi terbaik tetap diperoleh dengan hanya menggunakan algoritma SMO dalam melakukan diagnose penyakit kanker payudara yaitu sebesar 97,71%

Penelitian ini mengusulkan penggunaan seleksi fitur berbasis filter dan wrapper untuk diagnosa penyakit kanker payudara. Chi square dipilih sebagai algoritma seleksi fitur karena mampu menghasilkan akurasi yang lebih dalam proses klasifikasi [[3](#Yim971)]. Forward Feature Selection (FFS) dan Backward Feature Elimination (BFE) dipilih sebagai algoritma seleksi berbasis wrapper dikarenakan seringnya kedua algoritma ini digunakan.

Selanjutnya akan dijelaskan mengenai landasan teori di bab 2. Bab 3 menjelaskan metode penelitian. Bab 4 memaparkan hasil penelitian, dan bab 5 memberi kesimpulan dan saran untuk penelitian mendatang.

# Landasan Teori

## Algoritma Naïve Bayes

Naïve bayes merupakan algoritma yang cukup sederhana sehingga mampu melakukan proses klasifikasi dengan cepat [[4](#Tha09)]. Naïve bayes merupakan salah satu algoritma klasifikasi yang artinya memerlukan proses pelatihan dalam melakukan prediksi. Naïve bayes dihitung berdasarkan dari teorema Bayesian. Naïve Bayes menggunakan konsep probabilitas yang merupakan salah satu model statistik [[5](#Kus09)].

Model statistik merupakan salah satu model yang terpercaya sangat andal sebagai pendukung pengambilan keputusan. Konsep probabilitas merupakan salah satu bentuk model statistik. Salah satu metode yang menggunakan konsep probabilitas adalah Naive Bayesian Classification (NBC). Pada metode ini, semua atribut akan memberikan kontribusinya dalam pengambilan keputusan, dengan bobot atribut yang sama penting dan setiap atribut saling bebas satu sama lain. Apabila diberikan k atribut yang saling bebas (independence), nilai probabilitas dapat diberikan sebagai berikut.

 (1)

Jika atribut ke- bersifat diskret, maka  diestimasi sebagai frekuensi relatif dari sampel yang memiliki nilai  sebagai atribut ke  dalam kelas . Namun, jika atribut ke- bersifat kontinu, maka  diestimasi dengan fungsi densitas Gauss.

  (2)

## Seleksi Fitur Berbasis Filter

Dalam penelitian ini digunakan Chi Square (CHI) sebagai seleksi fitur berbasis filter. Seleksi fitur adalah proses menghilangkan beberapa fitur atau term yang kurang relevan untuk penentuan topik suatu dokumen. Terdapat 2 pembagian dalam seleksi fitur yaitu seleksi fitur supervised dan unsupervised. Seleksi fitur yang termasuk ke dalam kategori supervised adalah Chi-Square , Information Gain (IG) dan Mutual Information (MI), sedangkan seleksi fitur yang termasuk kedalam kategori unsupervised adalah Term Strength (TS), Term Contribution (TC), Entropy-based Ranking (En) dan document frequency (DF) [7]. Perbedaan antara keduanya adalah keberadaan informasi awal tentang kategori dari suatu data. Dalam pengklasifikasian sebuah data, CHI adalah salah satu supervised seleksi fitur yang mampu mengilangkan banyak fitur tanpa mengurangi tingkat akurasi.

## Seleksi Fitur Berbasis Wrapper

Dalam penelitian ini digunakan forward feature selection (FFS) dan backward feature elimination (BFE) sebagai seleksi fitur berbasis wrapper. FFS dan BFE mempunyai tujuan yang sama yaitu memilih atribut atau fitur yang dapat menyebabkan meningkatnya akurasi klasifikasi atau prediksi. BE diawali dengan menggunakan keseluruhan set peubah jelas, kemudian dilakukan eliminasi terhadap peubah tersebut satu persatu. Sedangkan FS dimulai dari himpunan kosong set peubah yang kemudian dilakukan penambahan peubah kedalam himpunan kosong tersebut satu persatu. Langkah-langkah algoritma FS dan BE dapat dilihat secara detail pada [[6](#Les08)].

# Metode Penelitian

Metode penelitian ini dilakukan dalam beberapa tahapan: studi literature, pengumpulan data, eksperimen, dan analisa hasil penelitian. Beberapa jurnal dikumpulkan untuk melakukan justifikasi terhadap pemilihan masalah dan solusi yang digunakan. Penelitian ini menggunakan dataset publik yang mudah digunakan. Sebagai hasil dari penelitian ini, didapatkan suatu kesimpulan yang mampu menjawab dari suatu masalah yang dinyatakan dalam penelitian ini. Gambar 1 memperlihatkan alur kerja metode yang diusulkan.

Dataset

Seleksi Fitur Berbasis Filter

Seleksi Fitur Berbasis Wrapper

Naïve Bayes

Gambar 1. Alur kerja metode yang diusulkan.

# Hasil Penelitian

## Data Penelitian

Penelitian ini menggunakan WDBC dataset yang merupakan dataset publik. Dataset tersebut merupakan dataset tentang pengidap kanker payudara yang dapat didownload di UCI dataset. Penjelasan detail dari tiap atribut tersebut dapat dilihat pada laman web UCI. Data tersebut terdiri dari 569 data (357 benign, 212 malignant) dan 30 atribut. Untuk setiap *nucleus* yang ada, akan dilakukan perhitungan sebagai berikut:

#### radius (mean of distances from center to points on

####  the perimeter)

#### texture (standard deviation of gray-scale values)

#### perimeter area

#### smoothness (local variation in radius lengths)

#### compactness (perimeter^2 / area - 1.0)

#### concavity (severity of concave portions of the contour)

#### concave points (number of concave portions of the contour)

#### symmetry

#### fractal dimension ("coastline approximation" - 1)

## Pengukuran Tingkat Keakuratan

Untuk mengetahui keberhasilan penelitian ini, kami menggunakan *confusion matrix* (Tabel 1) untuk mengetahui tingkat akurasi diagnosa penyakit kanker payudara dalam melakukan prediksi. Pengukuran tingkat akurasi ditunjukan pada persamaan (3).

Tabel 1: *Confusion Matrix*

|  |  |  |
| --- | --- | --- |
|  |  | Kelas Prediksi |
|  |  | Kelas 1 | Kelas 2 |
| Kelas Aktual | Kelas 1 | A | C |
| Kelas 2 | B | D |

  (3)

## Hasil Penelitian

Penelitian ini menggunakan software Yale (Rapidminer) dalam melakukan proses eksperimen. Dalam eksperimen ini, kami menguji penggunaan CHI, FFS dan BFE sebagai metode seleksi fitur dalam proses diagnosa penyakit kanker payudara yang dilakukan oleh algoritma Naïve Bayes. Keberhasilan metode seleksi fitur tersebut ditunjukan melalui tingkat akurasi algoritma Naïve Bayes.

Gambar 1 menunjukan perbandingan tingkat akurasi seleksi fitur FFS dan BFE dalam algoritma Naïve bayes. Dalam gambar tersebut nampak bahwa FFS lebih unggul dibandingkan dengan BFE dalam melakukan proses seleksi fitur. Penggunaan FFS dan BFE mampu meningkatkan kinerja Naïve Bayes yang hanya memiliki akurasi sebesar 94.17%.

Untuk meningkatkan kinerja FFS dan BFE dalam melakukan proses seleksi fitur, peneliti menambahkan penggunaan CHI. Hasilnya nampak pada gambar 3, 4, dan 5. Pada gambar tersebut tampak kinerja Naïve Bayes saat terjadi pengurangan fitur. Dalam melakukan seleksi fitur berbasis filter, peneliti menggunakan skala 20%, 40%, 60%, dan 80%. Ketika menggunakan prosentase fitur 20%, artinya fitur sebanyak 80% akan dihilangkan. Dengan perbandingan penggunaan *k*-fold validation, hasil akurasi terbaik didapat pada penggunaan CHI dan BFE pada 10-fold validation dengan tingkat akurasi 96,49%. Dibandingkan dengan penggunaan FFS saja (seleksi fitur berbasis wrapper), CHI dapat meningkatkan akurasi sebesar 0,17% (96,49% - 96,32%).

Gambar 2. Perbandingan tingkat akurasi penggunaan FFS dam BFE dalam berbagai k-fold validation.

Gambar 3. 10-fold validation

Gambar 4. 20-fold validation

Gambar 5. 30-fold validation

# Kesimpulan dan Saran

Penelitian ini bertujuan untuk melakukan evaluasi penggunaan selekfi fitur berbasis filter dan sebagai metode seleksi fitur pada algoritma naïve bayes. Penelitian ini menunjukan bahwa algoritma seleksi fitur tersebut mampu meningkatkan akurasi Naïve Bayes.

Berdasarkan hasil eksperimen yang dilakukan pada penelitian ini, bahwa metode yang diusulkan ini hanya mampu meningkatkan akurasi algoritma Naïve Bayes dalam diagnosa penyakit kanker payudara. Namun belum mampu meningkatkan akurasi pada penelitian sebelumnya, sehingga dalam penelitian mendatang dapat diusulkan suatu metode yang mampu meningkatkan akurasi pada penelitian sebelumnya.

##### Daftar Pustaka

x

|  |  |
| --- | --- |
| [1] | Gouda I. Salama, M.B.Abdelhalim, and Magdy Abd-elghany Zeid, "Breast Cancer Diagnosis on Three Different Datasets using Multi-Classifiers," *International Journal of Computer and Information Technology*, vol. 1, no. 1, September 2012. |
| [2] | Vaibhav Narayan Chunekar and Hemant P. Ambulgekar, "Approach of Neural Network to Diagnose Breast Cancer on three different Data Set," in *International Conference on Advances in Recent Technologies in Communication and Computing*, 2009. |
| [3] | Yiming Yang and Jan O.Pedersen, "A Comparative Study on Feature Selection in Text Categorization," in *Proceedings of ICML-97*, 1997. |
| [4] | Fadi Thabtah, Mohammad Ali H. Eljinini, Mannam Zamzeer, and Wa’el Musa Hadi, "Naïve Bayesian Based on Chi Square to Categorize Arabic Data," *Communications of the IBIMA*, vol. 10, no. 20, pp. 158-163, 2009. |
| [5] | Sri Kusumadewi, "Klasifikasi Status Gizi Menggunakan Naive Bayesian Classification," *CommIT*, vol. 3, no. 1, pp. 6 - 11, Mei 2009. |
| [6] | Iin Lesmanawati, "Pengklasifian nilai mutu ujian komprehensif mahasiswa departemen IPB menggunakan semi naive bayesian classifier," Institut Pertanian Bogor, Bogor, Skripsi 2008. |

x