PENGGABUNGAN ALGORITMA *BACKWARD ELIMINATION* DAN *NAIVE BAYES* UNTUK MENDIAGNOSIS PENYAKIT KANKER PAYUDARA

Laily Hermawanti
Jurusan Teknik Informatika, Universitas Sultan Fatah
Jl. Diponegoro 1A, Jogoloyo – Demak, Indonesia
Email: lailyhermawanti@gmail.com

Abstrak

Naive Bayes merupakan salah satu algoritma yang diusulkan oleh para peneliti data mining di bidang kesehatan misalnya penyakit kanker payudara. Penyakit kanker payudara merupakan salah satu penyakit berbahaya dan penyebab kematian di seluruh dunia. Maka dari itu, penyakit kanker payudara perlu didiagnosis. Algoritma yang digunakan dalam penelitian ini adalah penggabungan algoritma Backward Elimination dan Naive Bayes untuk meningkatkan akurasi dalam diagnosis penyakit kanker payudara. Penelitian ini menggunakan dataset kanker payudara yang diperoleh dari Wisconsin Breast Cancer (WBC) UCI Dataset Machine Learning Repository. Parameter-parameter yang digunakan pada data set Wisconsin Breast Cancer (WBC) adalah clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses dan class. Hasil penelitian ini, pada dataset kanker payudara, algoritma Naive Bayes memiliki nilai akurasi sebesar 96.14% +/- 2.13% dan nilai Area Under Curve (AUC) sebesar 0.978 +/- 0.017 yang termasuk dalam kategori klasifikasi sangat baik (excellent classification). Pada dataset kanker payudara, penggabungan algoritma Backward Elimination dan Naive Bayes memiliki nilai akurasi sebesar 97.00% +/- 2.56% dan nilai Area Under Curve (AUC) sebesar 0.979 +/- 0.022 yang termasuk dalam kategori klasifikasi sangat baik (excellent classification). Tingkat akurasi penggabungan algoritma Backward Elimination dan Naive Bayes lebih tinggi dari pada algoritma Naive Bayes dalam mendiagnosis penyakit kanker payudara.

Kata kunci : *backward elimination, naive bayes, penyakit kanker payudara*

PENDAHULUAN

Kanker payudara adalah kanker yang paling umum pada wanita dan penyebab utama kematian kanker di seluruh dunia (E. Technical and P. Series, 2006). Meskipun etiologi kanker payudara tidak diketahui, faktor risiko berbagai kemungkinan mempengaruhi perkembangan penyakit ini termasuk faktor genetik, hormonal, dan lingkungan. Selama beberapa dekade terakhir, risiko kanker payudara meningkat di negara-negara industri dan berkembang sebesar 1% -2% per tahun, tingkat kematian akibat kanker payudara menurun sedikit (E. Technical and P. Series, 2006). Parameter-parameter kanker payudara terdiri dari sebagai berikut (Frank, A. and Asuncion, 2011):

- *Clump thickness*: sel *benign* cenderung dikelompokkan dalam monolayers, sementara sel-sel kanker sering dikelompokkan dalam multilayers.
- *Uniformity of cell size*: sel-sel kanker mempunyai ukuran bervariasi.
- *Uniformity of cell shape*: sel-sel kanker mempunyai bentuk bervariasi.
- *Marginal adhesion*: sel-sel normal cenderung tetap bersama-sama.
- *Single epithelial cell size*: sel-sel epitel yang signifikan diperbesar menjadi sel *malignant.*
- *Bare nuclei*: adalah istilah yang digunakan untuk inti (*nuclei*) yang tidak dikelilingi oleh cytoplasm (seluruh sel). Biasanya terlihat di benign.
- *Bland Chromatin*: inti "tekstur" seragam yang dilihat dalam sel *benign*. Dalam sel-sel kanker chromatin cenderung lebih kasar.
- *Normal nucleoli*: *nucleoli* adalah struktur kecil yang terlihat dalam inti atom. Pada sel-sel normal nucleolus biasanya sangat kecil jika terlihat sama sekali. Dalam sel-sel *cancer nucleoli* menjadi lebih menonjol.
- *Mitoses* : pembelahan satu sel menjadi dua sel.
- *Class* : kelas.

Data mining dapat diaplikasikan di bidang kesehatan misalnya mendiagnosis penyakit kanker payudara, penyakit jantung, penyakit diabetes dan lain-lain (Larose, 2005). Terdapat beberapa metode dalam mendiagnosis penyakit

kanker payudara misalnya *Naïve Bayes* (Wu, J. dan Cai, Z., (2011), *K-Nearest Neighbor* (H. A. Fayed dan A. F. Atiya, 2009), dan lain-lain.

Penelitian yang dilakukan oleh J. Wu dan Z. Cai menggunakan banyak metode efektif untuk meningkatkan performa *Naive Bayes.* Pembahasan ini mengevaluasi performa konfigurasi baru (DE-WNB) pada 36 UCI seluruh standar *data set* dalam sistem Weka. Hasil eksperimen menunjukkan akurasi klasifikasi algoritma baru DE-WNB lebih tinggi dari algoritma lain yang digunakan untuk membandingkan. Algoritma *Differential Evolution Weighted Naïve Bayes* (DE-WNB) menghasilkan keakuratan sebesar 73.09%±7.51% untuk *dataset* kanker payudara *wisconsin*. Algoritma *Naïve Bayes* (NB) menghasilkan keakuratan sebesar 72.94%±7.71%. Algoritma *Gain Ratio-Weighted Naïve Bayes* (GR-WNB) menghasilkan keakuratan sebesar 70.30%±1.37%. Algoritma *Correlation-based Feature Selection-Weighted Naïve Bayes* (CFS-WNB) menghasilkan keakuratan sebesar 71.73%±7.40%. Algoritma *Mutual Information-Weighted Naïve Bayes* (MI-WNB) menghasilkan keakuratan sebesar 70.30%±1.37%. Algoritma *Tree-Weighted Naïve Bayes* (*Tree*-WNB) menghasilkan keakuratan sebesar 72.39%±7.47% (Wu, J. dan Cai, Z., 2011),.

Penelitian yang dilakukan oleh H. A. Fayed dan A. F. Atiya menggunakan metode *K-Nearest Neighbor* (KNN). Penelitian ini menggunakan *data set* dari *UCI machine learning depository*. Permasalahan KNN adalah komputasi dan penyimpanan. KNN memerlukan penyimpanan seluruh *training set* yang menjadi penyimpanan untuk *data set* yang besar dan waktu komputasi yang lama pada tahap klasifikasi. Penelitian ini mengusulkan algoritma kondensasi baru. Algoritma baru tersebut adalah *Template Reduction for* KNN (TRKNN). Pendekatan TRKNN yang diusulkan untuk mengurangi ukuran *template set*. Metode TRKNN mempunyai kelebihan yaitu implementasinya sederhana dan komputasinya cepat. Algoritma TRKNN dengan menghasilkan akurasi sebesar 95% (H. A. Fayed dan A. F. Atiya, 2009).

Dari penelitian-penelitian yang pernah dilakukan tentang diagnosis penyakit kanker payudara terutama yang menggunakan algoritma *Naive Bayes*, akurasinya belum tinggi. Kelebihan-kelebihan spesifik model penggabungan algoritma *Naive Bayes* dan *Backward Elimination* pada penyakit kanker payudara yang akan diteliti dibanding teknik-teknik diagnosis lain yaitu *Backward Elimination* dapat mereduksi ukuran *data set* sehingga dapat meningkatkan akurasi pada *Backward Elimination* (Han, J. and Kamber, 2006). Maka dari itu, penelitian ini menggunakan penggabungan algoritma *Backward Elimination* dan *Naive Bayes* untuk mendiagnosis penyakit kanker payudara sehingga dapat meningkatkan akurasi dibandingkan dengan penelitian-penelitian sebelumnya.

METODE

Penelitian ini menggunakan proses *Cross-Standard Industry-Data Mining* (CRISP-DM) dengan tahap-tahap penelitian meliputi pemahaman bisnis, pemahaman data, pengolahan data, pemodelan dan evaluasi (Larose, 2005).

Tahap Pemahaman Bisnis

Penelitian ini dilakukan untuk menerapkan penggabungan algoritma *Backward Elimination* dan *Naive Bayes* untuk meningkatkan akurasi dalam mendiagnosis penyakit kanker payudara.

Tahap Pemahaman Data

Penelitian ini mengambil *dataset* kanker payudara dari *UCI Machine Learning* (Frank, A. and Asuncion, 2011).

Tahap Pengolahan Data

Teknik-teknik pengolahan data awal (data *pre-processing*) yang digunakan pada penelitian ini adalah (Han, J. and Kamber, 2006) :

1. Data *cleaning* dapat digunakan untuk data yang *missing value*. Karena ditemukan adanya data yang terlewat tidak terisi (*missing value*) pada data. Pengolahan data awal dilakukan untuk mengisi nilai yang *missing value* dengan pekerjaan *replace missing value* dilakukan.
2. Data *reduction* digunakan untuk menghasilkan *data set* yang volumenya lebih kecil. Salah satu strategi *data reduction* yang digunakan pada penelitian ini adalah *attribute subset selection. Attribute subset selection* digunakan untuk mereduksi ukuran *data set* dengan menghilangkan atribut-atribut yang tidak relevan atau *redudant*. Salah satu teknik

attribute subset selection yang digunakan pada penelitian ini adalah *Backward Elimination.*

Tahap Pemodelan

Model yang digunakan dalam tahap ini menggunakan penggabungan algoritma *Backward Elimination* dan *Naive Bayes*.

Algoritma *Naive Bayes*

Algoritma Naive Bayes merupakan pengklafikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class (Han, J. and Kamber, 2006). Klasifikasi Bayesian mempunyai bentuk umum (Han, J. and Kamber, 2006) :

$$P(X|H) = \frac{P(H|X)P(H)}{P(X)} \quad (1)$$

Keterangan :

X : data dengan class yang belum diketahui
H : hipotesis data x merupakan suatu class
P(H|X) : probabilitas hipotesis H berdasarkan kondisi X (posteriori probability)
P(H) : probabilitas hipotesis H (prior probability)
P(X|H) : probabilitas X berdasar kondisi pada hipotesis H
P(X) : probabilitas dari X

Langkah-langkah algoritma Naive Bayes adalah (Han, J. and Kamber, 2006) :

1. Membolehkan training set pada *tuple-tuple* dan mereka digabungkan menjadi label class. Biasanya, setiap *tuple* ditunjukkan sebagai vektor atribut *n-dimensional*, X = (x_1, x_2,..., x_n), menggambarkan ukuran *n* pada *tuple* dari atribut-atribut n, masing-masing A_1, A_2,..., A_n.
2. Misalkan ada *classes* m, C_1, C_2,..., C_n. Memberikan *tuple*, X, pengklasifikasi akan meramalkan X selama class pada *posteori probability* tertinggi, dinamakan X. Pengklasifikasi *Naive Bayes* meramalkan tuple X untuk *class* C_i jika dan hanya jika
   $P(C_i|X) > P(C_j|X)$ untuk $1 \leq j \leq m, j \neq i$.
   Class C_i untuk $P(C_i|X)$ diperbesar yang dinamakan *maximum posteori probability*. Dengan teori Bayes (Persamaan 1).
   $P(C_i|X) = (P(X|C_i)\ P(C_i)) : P(X)$ ..... (2)
3. Misalkan P(X) adalah konstan untuk semua class, hanya $P(X|C_i)\ P(C_i)$ perlu diperbesar. Jika *class prior probabilty* tidak diketahui maka itu biasanya dianggap sebagai *class* yang sama, maka, $P(C_1) = P(C_2) = ... = P(C_m)$. Maka dari itu, memperbesar $P(X|C_i)\ P(C_i)$. Catatan bahwa *class prior probabilty* diestimasi sebagai $P(C_i) = |C_{i,D}| / |D|$, dimana $|C_{i,D}|$ adalah nomor *training tuples* pada *class* C_i dalam D.
4. Memberikan *data set* dengan atribut-atribut, itu akan digunakan untuk menghitung ($P(X|C_i)$. Untuk mengurangi komputasi dengan cara mengevaluasi ($P(X|C_i)$, lalu *Naive* pada *class conditional independence* dibuat. Ini menunjukkan nilai atribut-atribut secara kondisional independen merupakan salah satu yang diberikan pada label *class tuple* (contohnya, tidak ada hubungan di antara atribut-atribut).
5. Untuk memprediksi label *class* X, $P(X|C_i)P(C_i)$ dievaluasi untuk setiap *class* C_i. Pengklasifikasi memprediksi *class* label pada *tuple* X adalah *class* C_i jika dan hanya jika
   $P(X|C_i)P(C_i) > P(X|C_j)P(C_j)$ untuk $1 \leq j \leq m, j \neq i$ ... (3)
   Pada sisi lain, diprediksi sebagai label *class* adalah *class* C_i untuk $P(X|C_i)P(C_i)$ adalah maksimum.

Algoritma *Backward Elimination*

Backward Elimination menghilangkan atribut-atribut yang tidak relevan (Han, J. and Kamber, 2006). Algoritma *Backward Elimination* didasarkan pada model regresi linear (Noori, R., Karbassi, 2011). Langkah-langkah *Backward Elimination* adalah :

1. Mulai semua variabel pada model F-statistik parsial dihitung setiap variabel pada model.
2. Menentukan variabel dengan F-statistik parsial terkecil dan menguji Fmin.
3. Jika Fmin tidak signifikan, dalam kasus ini, variabel dihilangkan dari model.
4. Menentukan variabel dengan F-statistik parsial .
5. Pada sisi lain, variabel dengan F-statistik terkecil adalah variabel indikator. Bagaimanapun, p-value diasosiasikan dengan Fmin tidak cukup membenarkan model yang tidak inklusi (*noninclusion*) menurut kriteria (lebih dari bit). Maka dari itu, prosedur menghasilkan dan melaporkan model sebagai berikut :
   $y = \beta_0 + \beta_1$(*single epithelial cell size*) + β_2(*normal nucleoli*) + β_3(*marginal adhesion*) + ε
6. Menghitung F-test parsial.

Tahap Evaluasi

Evaluasi dan validasi pada penelitian ini menggunakan *confusion matrix* (*accuracy*) dan ROC *Curve*.

HASIL DAN DISKUSI

Akurasi *dataset* kanker payudara dapat dilihat pada Tabel 1. *Area Under Curve* (AUC) *dataset* kanker payudara dapat dilihat pada Tabel 2. Pada tabel 1, algoritma *Naive Bayes* menggunakan *dataset* kanker payudara menghasilkan akurasi sebesar 96.14% +/- 2.13%, sedangkan algoritma *Backward Elimination- Naive Bayes* menghasilkan akurasi sebesar 97.00% +/- 2.56% sehingga mengalami peningkatan akurasi. Hasil pada tabel 1, akurasi metode *Backward Elimination- Naive Bayes* lebih tinggi dari algoritma *Naive Bayes*.

Tabel 1. Akurasi *Dataset* Kanker Payudara

Algoritma	Akurasi (%)
Naive Bayes	96.14% +/- 2.13%
Backward Elimination – Naive Bayes	97.00% +/- 2.56%

Tabel 2. *Area Under Curve* (AUC) *Dataset* Kanker Payudara

Algoritma	*Area Under Curve* (AUC)
Naive Bayes	0.978 +/- 0.017
Backward Elimination – Naive Bayes	0.979 +/- 0.022

Area Under Curve (AUC) *dataset* kanker payudara dapat dilihat pada Tabel 2. Pada tabel 2, algoritma *Naive Bayes* menggunakan *dataset* kanker payudara menghasilkan *Area Under Curve* (AUC) sebesar 0.978 +/- 0.017 yang termasuk dalam kategori klasifikasi sangat baik *(excellent classification)*. Algoritma *Backward Elimination- Naive Bayes* menghasilkan AUC sebesar 0.979 +/- 0.022 yang termasuk dalam kategori "klasifikasi sangat baik (*excellent classification*)".

Hasil menunjukkan metode *Backward Elimination-Naive Bayes* dapat mencapai akurasi yang tinggi dalam mendiagnosis penyakit kanker payudara. Percobaan ini dilakukan untuk menunjukkan peningkatan akurasi dari algoritma *Naive Bayes* menjadi *Backward Elimination- Naive Bayes*.

KESIMPULAN

Tingkat akurasi penggabungan algoritma *Backward Elimination* dan *Naive Bayes* lebih tinggi dari pada algoritma *Naive Bayes* dalam mendiagnosis penyakit kanker payudara. Penelitian ini menunjukkan penggabungan algoritma *Backward Elimination* dan *K-Nearest Neighbor* merupakan salah satu algoritma yang tepat dalam mendiagnosis penyakit kanker payudara.

DAFTAR PUSTAKA

E. Technical and P. Series, (2006), *Guidelines for management of breast cancer*, World Health Organization. 2006.

Larose, D.T., (2005), *Discovering Knowledge in Data: An Introduction to Data Mining*, United States of America: John Wiley & Sons, Inc

Wu, J. and Cai, Z., (2011), *Attribute Weighting via Differential Evolution Algorithm for Attribute Weighted Naive Bayes* (WNB), vol. 5, pp. 1672–1679.

H. A. Fayed and A. F. Atiya, (2009), *A Novel Template Reduction Approach for the -Nearest Neighbor Method*, vol. 20, no. 5, pp. 890–896.

Han, J. and Kamber, (2006), *Data Mining Concept dan Techniques, 2nd ed,* United States of America: Diane Cerra.

Noori, R., Karbassi, A.R., A. Moghaddamnia, Han, D., Zokaei-ashtiani, M.H., and Farokhnia (2011), A. *Assessment of input variables determination on the SVM model performance using PCA , Gamma test , and forward selection techniques for monthly stream flow prediction, Journal of Hydrology*, vol. 401, no. 3–4, pp. 177–189.

Frank, A. and Asuncion, (2011), A. *UCI Machine Learning Repository*. http://archive.ics.uci.edu/ml/datasets.html, Irvine, CA: University of California, School of Information and Computer Science.