MENAKAR TINGKAT AKURASI SUPPORT VECTOR MACHINE STUDY KASUS KANKER PAYUDARA

Moh. Yamin Darsyah

[1]Program Studi S1 Statistika Universitas Muhammadiyah Semarang, Jl. Kedung Mundu Raya no 18 Semarang

Email: mydarsyah@yahoo.com

ABSTRAK

Kanker payudara adalah salah satu jenis kanker yang paling banyak menyerang kaum wanita. Menurut WHO 8-9% wanita akan mengalami kanker payudara. Pada tahun 2000 yang lalu WHO memperkirakan 1,2 juta wanita terdiagnosis kanker payudara dan lebih dari 700.000 meninggal dunia (WHO, 2005). Di Indonesia, pada tahun 2005 kanker payudara menduduki peringkat kedua setelah kanker leher rahim diantara kanker yang menyerang wanita Indonesia. Kanker ini sering menyebabkan kematian jika penangannya terlambat. Oleh karena itu, deteksi dini penyakit kanker payudara sangat diperlukan. Dewasa ini, penggunaan *machine learning* untuk diagnosis atau *prognosis* suatu penyakit telah banyak dilakukan. Dalam penelitian ini, ada metode yang terkenal dalam *machine learning* yaitu *Support Vector Machine (SVM)* digunakan untuk analisis diagnosis dan *prognosis* kanker payudara. SVM merupakan salah satu *machine learning* yang mempunyai beberapa kelebihan, diantaranya bisa memodelkan dan mengklasifikasikan hubungan antar variabel tanpa perlu asumsi yang ketat, efisien, dan interpretasinya mudah. SVM dalam mengklasifikasikan kategori penyakit penyakit kanker payudara akan bandingkan dengan metode statistika lainnya yaitu Regresi Logistik dan CART . Selanjutnya masing- masing hasil metode klasifikasi tersebut akan bandingkan dengan hasil dugaan K-Mean dan Kernel K-Mean Clustering. Maka dapat simpulkan seberapa akurat dan efisien antara SVM, Regresi Logistik, dan CART dalam ketepatan akurasi.

Kata Kunci: Kanker Payudara, SVM, Regresi Logistik, CART, K- Mean dan Kernel K-Mean

PENDAHULUAN

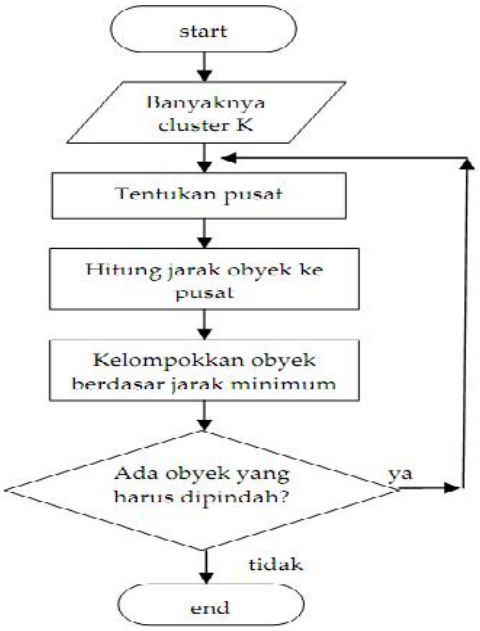
Penyebab pasti kanker payudara tidak diketahui. Meskipun demikian, riset mengidentifikasi sejumlah faktor yang dapat meningkatkan risiko pada individu tertentu, yang meliputi: keluarga yang memiliki riwayat penyakit serupa, usia yang makin bertambah, tidak memiliki anak, kehamilan pertama pada usia di atas 30 tahun, periode menstruasi yang lebih lama (menstruasi pertama lebih awal atau menopause lebih lambat), faktor hormonal (baik estrogen maupun androgen). *Support Vector Machine* (SVM) pertama kali dikenalkan oleh Vapnik pada tahun 1995 dan merupakan salah satu metode yang baik untuk klasifikasi. Saat ini SVM telah sukses diterapkan dalam berbagai permasalahan seperti credit scoring

Pembentukan cluster merupakan salah satu teknik yang digunakan dalam mengekstrak pola kecenderungan sua tu data. Clustering memegang peranan penting dalam aplikasi data mining, misalnya eksplorasi data ilmu pengetahuan, pengaksesan informasi dan text mining, aplikasi basis data spasial, dan analisis web.

Metode nonhierarki yang paling popular adalah metode *K-means*. Metode ini merupakan metode pengelompokan

yang bertujuan mengelompokkan objek sedemikian hingga jarak tiap - tiap objek ke pusat kelompok di dalam satu kelompok adalah minimum dimana jumlah kelompok dalam metode *K-Means cluster* ditentukan terlebih dahulu.

Berikut diagram alir dari algoritma K-Means Clustering

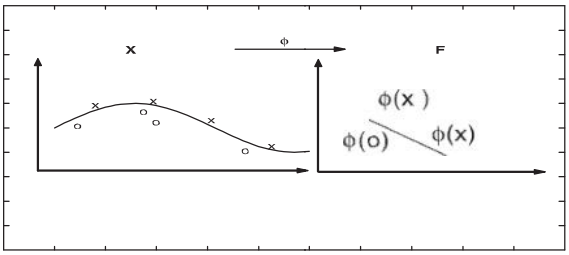


Gambar 1. algoritma K-Means Clustering

Kernel K-Means merupakan pengembangan dari algoritma K-Means dimana digunakan metode Kernel. Hal ini dilakukan untuk meningkatkan akurasi hasil pengelompokan. Kernel K-Means diharapkan dapat memisahkan data dengan lebih baik utamanya untuk data *overlap* atau *outlier* bisa menjadi linier di ruang dimensi baru (Santosa, 2007).

Support Vector Machine Kernel

Pada umumnya, dalam permasalahan nyata, jarang ditemukan data *linear separable*. Sehingga fungsi Kernel digunakan dalam Support Vector Machine untuk mengatasi data non-linier. Dengan memasukkan fungsi Kernel, maka problem data non-linier menjadi linier dalam space baru seperti tampak pada ilustrasi berikut.



Gambar 2. Ilustrasi Problem Non-Linier menjadi Linier dengan Kernel SVM

Secara matematis, beberapa fungsi Kernel dijelaskan sebagai berikut.

1. Kernel Linier: $x^T x$,
2. Kernel Polynomial: $\left(x^T x_i + 1\right)^p$,
3. Kernel Radial Basis Function (RBF): $\exp\left(-\frac{1}{2\pm^2}\left\|x - x_i\right\|^2\right)$,
4. Kernel Tangent hyperbolic: $\tanh\left(x^T x_i + \right)$, dimana , $\in \Re$

Analisis regresi logistik biner (Hosmer,1989) adalah suatu regresi logistik antara variabel respon (y) dan variabel prediktor (x) dimana variabel y menghasilkan 2 kategori yaitu 0 dan 1. Sehingga variabel y mengikuti distribusi Bernoulli dengan fungsi probabilitasnya sebagai berikut

$f(y_i)=®(x_i)^{y_i}(1-®(x_i))^{1-y_i}$ dim ana y_i = 0,1 (4.1) adalah probabilitas kejadian yang diakibatkan oleh variabel prediktor(x)

Rata-ratanya adalah sebagai berikut :

$E(y_i) = 1\,[P(y=1)] + 0\,[P(y=0)] = P\,[y=1]$ dan probabilitas tersebut dinotasikan ® (x_i) yang menggambarkan ketergantungannya akan nilai variabel prediktor (x) sehingga

$E(y_i^2) = 1^2\,[®(x_i)] + 0^2\,[1-®(x_i)] = ®(x_i)$

Jadi varians dari variabel respon (y) adalah

$V(y_i) = E(y_i^2) - [E(y_i)]^2 = ®(x_i)\,[1-®(x_i)]$.

Untuk menyatakan rata-rata bersyarat dari y jika diberikan nilai x digunakan nilai $®(x)$ = *E(y/x)*. Sedangkan bentuk model regresi logistiknya dinyatakan sebagai :

$$®(x) = \frac{\exp(\ _0 + \ _1 x)}{1 + \exp(\ _0 + \ _1 x)}$$

Pada regresi logistik ini dapat disusun model yang terdiri dari banyak variabel prediktor dikenal sebagai model multivariabel. Model regresi logistik multivariabel dengan p variabel prediktor adalah:

$$®(x) = \frac{\exp(\ _0 + \ _1 x_1 + ... + \ _p x_p)}{1 + \exp(\ _0 + \ _1 x_1 + ... + \ _p x_p)}$$

Bila model persamaan di atas ditranformasi dengan tranformasi logit, maka didapatkan bentuk logit:

$$\hat{g}(x) = \ _0 + \ _1 x_1 + ... + \ _p x_p$$

Cart

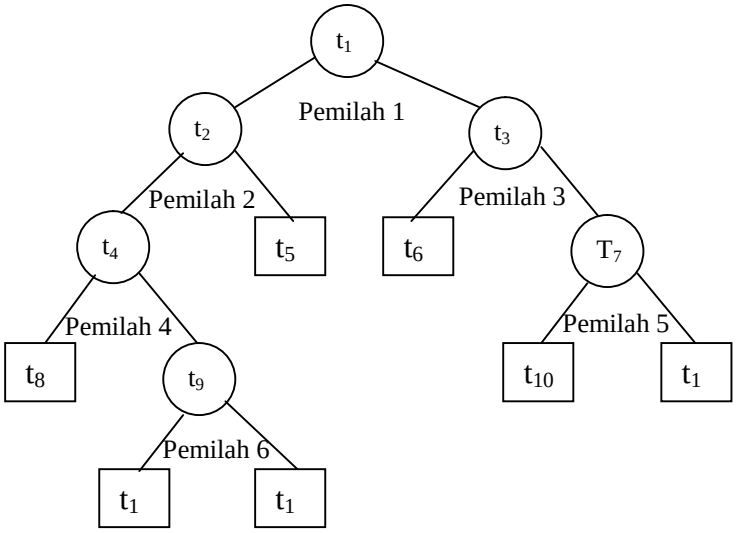
Merupakan salah satu Regresi Nonparametrik untuk memprediksikan dan mengklasifkasi suatu model (Friedman, 1991). Sifat- sifat CART antara lain: tidak memerlukan spesifikasi bentuk fungsional modelnya, infarian terhadap tranformasi monoton dari peubah bebasnya, tegar terhadap pengaruh pencilan, dapat menangani peubah kategorik maupun kontinu secara lebih baik, dapat menangani pengamatan dengan data hilang pada satu atau beberapa peubah bebasnya. Metode dalam CART dibagi 2:

a) Regresi pohon untuk pemodelan variabel respon kontinu;
b) Klasifikasi pohon untuk pemodelanvariabel respon kategorik.

Kriteria penyekatan menggunakan formula sebagai berikut:

$$\phi(s,t) = JKS(t) - [JKS(t_L) + JKS(t_R)]$$

Ilustrasi pohon klasifikasi tampak pada gambar berikut:



Gambar 1. Ilustrasi Pohon Klasifikasi CART

Algoritma CART secara umum melalui tiga tahapan yaitu pembentukan pohon klasifikasi, pemangkasan pohon klasifikasi, dan penentuan pohon klasifikasi optimum sehingga diperoleh hasil klasifikasi akhir.

METODE PENELITIAN

Data yang digunakan dalam penelitian diambil dari data pasien kanker payudara UCI Machine learning Wicoxsin University.

Data untuk diagnosis kanker payudara adalah data pasien yang melakukan skrining mamografi dan melakukan biopsy. Variabel penelitian yang digunakan pada penelitian ini terdiri dari variabel respon (*y*) dan variabel prediktor (*x*). Variabel respon dalam penelitian ini dibagi menjadi 2 kategori, yaitu :

- kanker jinak (-1)
- kanker ganas (1)

Berdasarkan literatur , (x1-x9) variable predictor sebagai berikut :

I0	Impedivity (ohm) at zero frequency
PA500	phase angle at 500 KHz
HFS	high-frequency slope of phase angle
DA	impedance distance between spectral ends
AREA	area under

	spectrum
A/DA	area normalized by DA
MAX IP	maximum of the spectrum
DR	distance between I0 and real part of the maximum frequency point
P	length of the spectral curve

Pada penelitian ini akan membandingkan tiga metode analisis klasifikasi yaitu SVM, Regresi Logistik Biner, dan CART. Adapun Metode Kernel yang digunakan ialah Kernel RBF, dan Polynimial. Selain dengan menggunakan variabel dependen (*Y*) dari data asli, klasifikasi juga dicoba menggunakan variabel dependen (*Y*) hasil dugaan metode K-means clustering dan hasil duguan Kernel K-means clustering.

Langkah selanjutnya sebagai berikut:

1. Melakukan analisis pengelompokkan/clustering menggunakan data variabel independen (*X*) dengan metode K-means clustering.
   a. Jumlah kelompok *K* ditentukan sebanyak 2 kelompok.
   b. Hasil pengelompokan K-means clustering tiap objek kemudian dibandingkan dengan variabel dependen (*Y*).
   c. Pengelompokan objek yang sama dengan variabel dependen (*Y*) kemudian dihitung persentasenya untuk dilihat ketepatan pengelompokannya.
   d. Hasil pengelompokan K-means clustering selanjutnya digunakan sebagai variabel dependen (*Y*) taksiran untuk selanjutnya dilakukan analisis klasifikasi selanjutnya.
2. Melakukan analisis pengelompokkan/clustering menggunakan data variabel dependen (*X*) dengan metode kernel K-means clustering.
   a. Jumlah kelompok *K* ditentukan sebanyak 2 kelompok.
   b. Variasi metode kernel yang digunakan untuk kernel K-means clustering ialah kernel RBF, dan polynomial.
   c. Parameter kernel yang dicobakan ialah 1 dengan *maximum iteration* 10.
   d. Hasil pengelompokan kernel K-means clustering untuk masing-masing variasi metode kemudian dibandingkan dengan variabel dependen (*Y*).
   e. Pengelompokan objek yang sama dengan variabel dependen (*Y*) kemudian dihitung persentasenya untuk dilihat ketepatan pengelompokannya.
3. Melakukan analisis klasifikasi dengan regresi logistik biner.
   a. Analisis klasifikasi regresi logistik biner dilakukan menggunakan variabel dependen (*Y*) dari data asli, *Y* dugaan hasil K-means, dan *Y* dugaan hasil kernel K-means terbaik.
   b. Penentuan metode terbaik digunakan backward elimination Wald.
   c. Mengamati ketepatan klasifikasi hasil analisis dan kemudian membandingkannya antar variasi variabel dependen *Y* yang digunakan.
4. Melakukan analisis klasifikasi dengan SVM Kernel.
   a. Analisis klasifikasi SVM kernel dilakukan menggunakan variabel dependen (*Y*) dari data asli, *Y* dugaan hasil K-means, dan *Y* dugaan hasil kernel K-means terbaik.

b. Metode kernel yang digunakan ialah kernel RBF, dan polynomial.
c. nilai C ditentukan bernilai 10.
d. Mengamati ketepatan klasifikasi hasil analisis dan membandingkannya antar variasi parameter kernel dan antar variasi variabel dependen *Y* yang digunakan.

5. Melakukan analisis klasifikasi dengan CART.
   a. Analisis klasifikasi CART dilakukan menggunakan variabel dependen (*Y*) dari data asli, *Y* dugaan hasil K-means, dan *Y* dugaan hasil kernel K-means terbaik.
   b. V-fold cross validation ditentukan bernilai 10.
   c. Mengamati ketepatan klasifikasi hasil analisis dan kemudian membandingkannya antar variasi variabel dependen *Y* yang digunakan.
6. Menghitung ketepatan klasifikasi untuk tiap metode klasifikasi dari berbagai variabel dependen *Y* .
7. Membuat kesimpulan hasil penelitian tentang metode klasifikasi terbaik.

HASIL DAN PEMBAHASAN

Variabel independen (*X*) digunakan untuk analisis clustering guna mendapatkan taksiran variabel dependen (*Y*) dimana selanjutnya digunakan untuk analisis klasifikasi. Hasil pengelompokan dengan metode K-means clustering setelah dicocokkan dengan variabel dependen (*Y*) asli diperoleh ketepatan pengelompokan sebesar 59,6%. Hasil ini menunjukkan hasil yang kurang baik karena kecilnya ketepatan pengelompokan. Kemudian, hasil pengelompokan K-means ini digunakan sebagai taksiran variabel dependen (*Y*) untuk selanjutnya dilakukan analisis klasifikasi. Taksiran variabel dependen (*Y*) juga didapatkan dari hasil analisis kernel K-means clustering yang di cocokan dengan (y) data aslinya, Ketepatan pengelompokan kernel K-means untuk masing-masing variasi kernel Polinomial dan RBF sebesar 71,1% dan 73,6%. Jadi ketepatan pengelompokan yang paling tinggi akurasinya dengan metode Kernel K-Mean.

Dari hasil pengelompokan K-means dan kernel K-means clustering, maka ada tiga jenis variabel dependen (*Y*) yaitu variabel dependen (*Y*) data asli, variabel dependen (*Y*) hasil pengelompokan K-means, dan variabel dependen (*Y*) hasil pengelompokan kernel K-means. Kemudian dari tiga variabel dependen (*Y*) tersebut dengan variabel independen (*X*) asli, dilakukan analisis klasifikasi dengan metode SVM, Regresi logistik biner, dan CART.

Berikut hasil output perbandingan ketepatan akurasi klasifikasi Y :

1. Support Vector Machine

Tabel 1. Support Vector Machine

	Y asli	Y K-Mean	Y Kernel K- Mean
RBF	99%	84,4%	100%
Poly	96%	83,2%	99%

Dari tabel diatas dapat disimpulkan bahwa SVM dengan Y Kernel K- Mean memberikan akurasi 100% dalam klasifikasi, secara keseluruhan metode SVM memberikan hasil yang terbaik mungkin ini yang menjadikan SVM mendapat perhatian serius dari para peneliti karena bisa mencari hyperpline yang terbaik.

2. Analisa Regresi logistik

Tabel 2. Regresi logistik

Y asli	Y K-Mean	Y Kernel K-Mean
61,7%	66,8%	73,8%

Dari tabel diatas dengan menggunakan analisa regresi logistik akurasi dari Y asli memberikan nilai terendah sebesar 61,7% dengan menggunakan metode backward. Dan akurasi tertinggi dengan menggunakan Y Kernel K-Mean sebesar 73,8%

3. CART

Tabel 3. CART

Y asli	Y K-Mean	Y Kernel K-Mean
72,7%	78,8%	81,8%

Dari tabel diatas terlihat bahwa CART dengan Y Kernel K-Mean memberikan hasil akurasi yang tinggi dan akurasi terendah Y asli.

Berikut disajikan dalam tabel untuk ketiga metode tersebut tampak sebagi berikut:

Tabel 4. Perbandingan Tiga Metode

Metode Klasifikasi	**Y asli**	**Y K-means**	**Y kernel K-means**
SVM (RBF)	99%	84,4%	100%
Regresi Logistik	61,7%	66,8%	73,8%
CART	72,7%	78,8%	81,8%

KESIMPULAN

Hasil penelitian diatas menunjukan bahwa dari beberapa metode klasifikasi, SVM terlihat mempunyai tingkat akurasi lebih tinggi dalam metode klasifikasi.

DAFTAR PUSTAKA

Abe, S. 2010. *Support vector machines for pattern classification*. 2nd ed. New York: Springer-Verlag.

Abonyi, J., Szeifert, F., Supervised fuzzy clustering for the identification of fuzzy classifiers, *Patern. Recogn*., Lett. 24, pp. 2195-2207, 2003.

Altman, D.G., 1991. *Practical Statistics for Medical Research*, Chapman & Hall, London.

Anderson, S. Dkk., 1980. *Statistical Method for Comparative Studies Techniques for Bias Reduction*, John Wiley and Sons. Canada.

Bennet, K.P., and Blue, J.A., A Support Vector Machine Approach to Decision Tree, Math. Report, vol. 97-100, Rensselaer Polytechnic Institute, 1997.

Dinas Kesehatan Nasional. 2007. Data penderita kanker payudara di Indonesia. http://www.depkes.go.id/index.php/berita/press-release/1060-jika-tidak-dikendalikan-26-juta-orang-di-dunia-menderita-kanker-.html. (31 januari 2011)

Djatmiko, dkk .2009. *Breast Physician* Course. Surabaya: Klinik Onkologi Surabay.

Kardinah. 2002. *Penatalaksanaan Kanker Payudara Terkini oleh Penanggulangan & Pelayanan Kanker Payudara Terpadu Paripurna R.S. Kanker Dharmais*. Jakarta: Pustaka Populer Obor.