 ISSN 2089-5490

ANALISIS ALGORITHMS SUPPORT VECTOR MACHINE DENGAN NAIVE BAYES KERNEL PADA KLASIFIKASI DATA

Hanna Willa Dhany[1], Fahmi Izhari[2]
Fakultas Sains dan Teknologi
Universitas Pembangunan Pancabudi Medan, Medan, Indonesia
E-mail: hdhany@dosen.pancabudi.ac.id,, fahmi_izhari@dosen.pancabudi.ac.id

Abstrak—Penelitian ini membahas tentang Support Vector Machine (SVM) dan Naive Bayes dalam penambangan data. Banyak peneliti melakukan dan mengembangkan metode untuk meningkatkan akurasi dan klasifikasi data dalam hasil yang baik. Penelitian ini dilakukan dengan melakukan percobaan pada jenis bunga. Dalam penelitian ini, disimpulkan bahwa kinerja Naïve Bayes lebih baik daripada Support Vector Machine, Naïve Bayes memiliki hasil yang sangat baik yang berjanji untuk membantu mengklasifikasikan nilai-nilai terbaik untuk mendapatkan pengelompokan data. Support Vector Machine lebih besar dengan nilai 89.66% dari Naïve bayes yang menghasilkan 89.29% yang memiliki selisih 0.37 % dari kedua algoritma tersebut. Sedangkan Class Recall yang dihasilkan Naïve Bayes mencapai 89.29% dan Support Vector Machine 92.86% dan dengan perbedaan nilai akurasi mencapai 3.22% dari perbandingan performa kedua algoritma tersebut.

Kata kunci—Support Vector Machine, Naïve Bayes, Akurasi, Klasifikasi.

I. PENDAHULUAN

Support Vector Machine (SVM) pertama kali diperkenalkan oleh Vapnik pada tahun 1992 sebagai rangkaian harmonis konsep-konsep unggulan dalam bidang pattern recognition. Sebagai salah satu metode pattern recognition,usia SVM terbilang masih relatif muda. Walaupun demikian, evaluasi kemampuannya dalam berbagai aplikasinyamenempatkannya sebagai state of the art dalam pattern recognition, dan dewasa ini merupakan salah satu tema yangberkembang dengan pesat.

Menurut (Lee, 2001) SVM memanfaatkan optimasi dengan quadratic programming, sehingga untuk data berdimensi tinggi dan data jumlah besar SVM menjadi kurang efisien. Oleh karena itu dikembangkan smoothing technique yang menggantikan plus function SVM dengan integral dari fungsi sigmoid neural network yang selanjutnya dikenal dengan Smooth Support Vector Machine (SSVM). Apabila dibandingkan dengan SSVM, SVM memiliki waktu running yang lebih lama dan akurasi yang lebih kecil daripada SSVM

Menurut (Rachman, 2011), (Huang, 2003) dan (Byvatov, 2003) Support Vector Machine memiliki tingkat akurasi yang lebih baik jika dibandingkan dengan metode regresi logistik, ANN, Naive Bayes, dan CART . Support Vector Machine merupakan metode berbasis machine learning yang sangat menjanjikan untuk dikembangkan karena memiliki performansi tinggi dan dapat diaplikasikan secara luas untuk klasifikasi dan estimasi.

Menurut penelitian (Honakan, 2018) klasifikasi dengan metode support vektor machine dapat akurasi tertinggi dengan kombinasi stopword, tokenizing, term frequency & chi-square 47,43 %. Sedangkan penelitian (Pratama, 2018) Support Vector Machine (SVM) mengklasifikasikan data menjadi 2 kelas menggunakan kernel Gaussian RBF dengan kombinasi nilai parameter $\lambda = 0{,}5$, konstanta $\gamma = 0{,}01$, dan ε (epsilon) = 0,001 itermax = 100, c = 1 dengan menggunakan data latih sebanyak 170 dataset. Penelitian ini menghasilkan rata-rata akurasi sebesar 80,55 %.

Penentuan data training dapat mempengaruhi hasil pengujian, karena pola data training tersebut akan dijadikan sebagai rule untuk menentukan kelas pada data testing. Sehingga besar atau kecilnya prosentase tingkat precision, recall, dan accuracy dipengaruhi juga oleh penentuan data training. (Ridwan, 2013)

Penerapan metode naïve bayes diharapkan mampu untuk memprediksi besarnya penggunaan listrik tiap rumah tangga agar lebih mudah mengatur penggunaan listrik. dari 60 data penggunaan listrik rumah tangga yang diuji dengan metode naïve bayes, maka diperoleh hasil persentase 78,3333% untuk keakuratan prediksi, di mana dari 60 data penggunaan listrik rumah tangga yang diuji terdapat 47 data penggunaan listrik rumah tangga yang berhasil diklasifikasikan dengan benar. (Saleh , 2015).

II. METODE PENELITIAN

A. Naïve Bayes

Naive Bayes Classifier (NBC) NBC merupakan salah satu algoritma dalam teknik data mining yang menerapkan teori Bayes dalam klasifikasi. Teorema keputusan Bayes adalah adalah pendekatan statistik yang fundamental dalam pengenalan pola (pattern recoginition). Naive bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara konditional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai

 ISSN 2089-5490

output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Dengan memasukkan Persamaan 1 ke Persamaan 2 akan diperoleh pendekatan yang digunakan dalam NBC. (Santosa, 2002)

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma mengunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas(Patil , 2013). Definisi lain mengatakan Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Bustami, 2013).

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu (Ridwan, 2013). Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Naive Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan(Pattekari, 2012).

Teorema Bayes dikemukakan oleh Thomas Bayes pada tahun 1763. Teorema Bayes digunakan untuk menghitung peluang atau probabilitas terjadinya suatu peristiwa berdasarkan pengaruh yang didapat dari hasil observasi. Perbedaan mendasar antara metode Bayesian dengan statistik pada umumnya adalah bahwa dalam Bayesian, parameter dianggap sebagai variabel random sedangkan dalam statistic klasik, parameter tidak diketahui dan tetap. Teorema Bayes, diambil dari nama Rev. Thomas Bayes, yang menggambarkan hubungan antara peluang bersyarat dari dua kejadian H dan x, dimana dapat dijelakan dalam sebuah rumus sebagai berikut (Kundu, 2011):

$$P(H \mid x) = \frac{P(x \mid H)\,P(H)}{P(x \mid H)P(H) + P(x \mid H)P(H)}$$

Atau

$$P(H \mid x) = \frac{P(x \mid H)\,P(H)}{P(x)}$$

Misalkan x adalah sampel data yang label kelasnya tidak diketahui, dan H adalah hipotesa, maka sampel data x termasuk dalam kelas khusus c. P(H/x) merupakan probabilitas yang menjelaskan bahwa hipotesa H berlaku dengan diberikannya sampel data hasil pengamatan x. P(H/x) adalah probabilitas posterior yang menggambarkan keyakinan pada hipotesa setelah x diberikan. Sebaliknya, P(H) adalah probabilitas H sebelumnya untuk sesuatu sampel, terlepas dari bagaimana bentuk data dalam sampel. Probabilitas posterior P(H/x) didasarkan pada lebih banyak informasi daripada probabilitas priori P(H). Teorema Bayes memberikan cara menghitung probabilitas posterior P(H/x) dengan menggunakan probabilitas P(H), P(x) dan P(x/H).

Metode Bayes merupakan pendekatan statistic untuk melakukan inferensi induksi pada persoalan klasifikasi. Pertama kali dibahas terlebih dahulu tentang konsep dasar dan definisi pada Teorema Bayes, kemudian menggunakan teorema ini untuk melakukan klasifikasi dalam *Data Mining*. Metode Bayes menggunakan propabilitas bersyarat sebagai dasarnya.

1. Prinsip Metode Bayes

a. Metode Bayes memberikan cara yang mendasar dalam memasukkan informasi eksternal ke dalam proses analisa data. Proses ini diawali dengan distribusi probabilitas yang sudah ada diberikan untuk himpunan data yang dianalisa (Albert, 2009). Prinsip metode Bayesian berdasarkan peluang bersyarat, sehingga dalam Bayesian mengenal dua istilah penting yaitu :
b. Prior yaitu distribusi dari parameter. Dalam menentukan prior dilakukan dengan tingkat ketersediaan informasi penelitian sebelumnya. Karena distribusi diberikan sebelum ada data yang dipertimbangkan, sehingga disebut distribusi priori.
c. Posterior adalah distribusi yang merupakan perkalian antara prior dengan fungsi likelihood. Hal ini juga merupakan perbedaan antara metode Bayesian dan statistic klasik dimana statistic klasik melakukan inferensia hanya berdasarkan fungsi *likelihood* sedangkan metode Bayesian

 ISSN 2089-5490

menggunakan distribusi posterior yeng merupakan perkalian antara fungsi *likelihood* dan prior. Himpunan data baru menjadikan distribusi priori ini menjadi distribusi posterior. Perubahan yang terjadi dari priori ke posterior merujuk pada Teorema Bayes.

Aplikasi Metode Bayes biasanya digunakan dalam beberapa kategori sebagai berikut:

a. Menentukan diagnosa suatu penyakit berdasarkan data-data gejala (sebagai contoh hipertensi atau sakit jantung).
b. Mengenali buah berdasarkan fitur-fitur buah seperti warna, bentuk, rasa dan lain-lain
c. Mengenali warna berdasarkan fitur indeks warna RGB
d. Mendeteksi warna kulit (skin detection) berdarkan fitur warna chrominant
e. Menentukan keputusan aksi (olahraga, art, psikologi) berdasarkan keadaan.
f. Menentukan jenis pakaian yang cocok untuk keadaan-keadaan tertentu (seperti cuaca, musim, temperatur, acara, waktu, tempat dan lain-lain)

2. Teknik Klasifikasi Metode Bayes

Beberapa teknik pengklasifikasian yang digunakan (Albert, 2009):

a. Decision tree classifier
b. Rule based classifier
c. Neural network
d. Naive bayes

Setiap teknik menggunakan alogoritma pembelajaran untuk mengindentifikasi model yang memberikan hubungan yang paling sesuai. Contoh dari teori bayesian adalah kasus pasien yang memiliki kesulitan dalam bernafas. Keputusan yang diambil adalah antara kasus pasien yang menderita asma atau pasien yang menderita kanker paru-paru (Bolstad, 2007).

a. Keputusan 1: menyatakan seseorang menderita kanker paru-paru walaupun sebenarnya gejala asma (cost: cukup tinggi, sehingga menakuti pasien dan membuat pasien menjalani pemeriksaan yang tidak perlu).
b. Keputusan 2: menyatakan seseorang asma walaupun sebenarnya kanker paru-paru (cost: sangat tinggi sehingga membuat pasien kehilangan kesempatan untuk mengobati kanker pada stadium awal ataupun akhir).

3. Keuntungan dan Kerugian Metode Bayes

Kerugian Metode Bayes antara lain:

a. Metode Bayes hanya bisa digunakan untuk persoalan klasifikasi dengan supervised learning dan data-data kategorikal.
b. Metode Bayes memerlukan pengetahuan awal untuk dapat mengambil suatu keputusan. Tingkat keberhasilan metode ini sangat tergantung pada pengetahuan awal yang diberikan.

Keuntungan Metode Bayes antara lain:

a. Interpolation: Metode bayes mempunyai pilihan mengenai seberapa besar waktu dan usaha yang dilakukan oleh manusia vs komputer.
b. Bahasa: Metode bayes mempunyai bahasa tersendiri untuk menetapkan hal prior dan posterior.
c. Intuisi: Melibatkan prior dan integrasi, dua aktivitas yang berguna secara luas.

Bayesian probability adalah teori terbaik dalam menghadapi masalah estimasi dan penarikan kesimpulan. Bayesian method dapat digunakan untuk penarikan kesimpulan pada kasus-kasus dengan multiple source of measurement yang tidak dapat ditangani oleh metode lain seperti model hierarki yang kompleks (Bolstad, 2007).

B. Smooth Support Vector Machine

Support Vector Machine (SVM) pertama kali diperkenalkan oleh Vapnik pada tahun 1992 sebagai rangkaian harmonis konsep-konsep unggulan dalam bidang pattern recognition. Sebagai salah satu metode pattern recognition, usia SVM terbilang masih relatif muda. Walaupun demikian, evaluasi kemampuannya dalam berbagai aplikasinya menempatkannya sebagai *state of the art* dalam *pattern recognition*. SVM adalah metode learning machine yang bekerja atas prinsip *Structural Risk Minimization* (SRM) dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space*. Tulisan ini membahas teori dasar SVM dan aplikasinya dalam informatika, khususnya pada analisa ekspresi gen yang diperoleh dari analisa

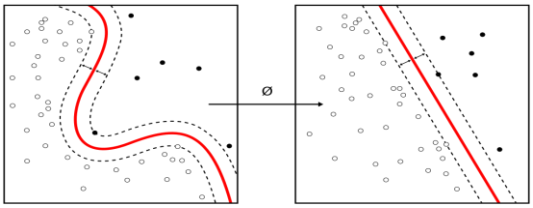
 ISSN 2089-5490

micro array. Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyper plane* terbaik yang berfungsi sebagai pemisah dua buah *class* pada *input space*.
Konsep dasar SVM sebenarnya merupakan kombinasi harmonis dari teori-teori komputasi yang telah ada puluhan tahun sebelumnya, seperti *margin hyperplane* (Duda & Hart tahun 1973, Cover tahun 1965, Vapnik 1964, dsb.), kernel diperkenalkan oleh Aronszajn tahun 1950, dan demikian juga dengan konsep-konsep pendukung yang lain. Akan tetapi hingga tahun 1992, belum pernah ada upaya merangkaikan komponen-komponen tersebut.

1. Karakteristik SVM
Karakteristik SVM sebagaimana telah dijelaskan pada bagian sebelumnya, dirangkumkan sebagai berikut:

a. Secara prinsip SVM adalah *linear classifier*
b. *Pattern recognition* dilakukan dengan mentransformasikan data pada *input space* ke ruang yang berdimensi lebih tinggi, dan optimisasi dilakukan pada ruang *vector* yang baru tersebut. Hal ini membedakan SVM dari solusi *pattern recognition* pada umumnya, yang melakukan optimisasi parameter pada ruang hasil transformasi yang berdimensi lebih rendah daripada dimensi *input space.*
c. Menerapkan strategi *Structural Risk Minimization* (SRM)
d. Prinsip kerja SVM pada dasarnya hanya mampu menangani klasifikasi dua class.

Support Vector Machine dapat dibayangkan sebagai permukaan yang mendefinisikan batas antara berbagai titik data yang mewakili contoh yang diplot dalam ruang multidimensi sesuai dengan fiturnya. Tujuan dari *SVM* adalah untuk membuat batas dasar atau yang disebut dengan *hyperline* yang mengarah pada partisi data yang homogeny diantara kedua sisi. Dengan cara ini, pembelajaran SVM yang menggabungkan aspek-aspek dari pembelajaran tetangga terdekat. Untuk permasalahan klasifikasi biner, *SVM* sangat cocok digunakan. Sebagai contoh pada gambar berikut ini.



Gambar 1 *Support Vector Machine*

III. IDENTIFIKASI MASALAH

Dari latar belakang masalah yang telah diuraikan, maka penulis mengambil rumusan masalah untuk pemrosesan data dibutuhkan beberapa metode untuk mendapatkan hasil yang lebih baik dan optimal. Dengan melakukan perbandingan dari metode yang digunakan sangat dibutuhkan untuk proses pengolahan data yang baik untuk menganalisa kinerja algoritma melalui perbandingan dari algoritma *Support Vector Machine* dan *Naive Bayes Kernel* dengan klasifikasi kelas yang berbeda dari sudut pandang *precision, recall & accuracy* dan *F-Measure.*

IV. HASIL DAN PEMBAHASAN

Dari percobaan terhadap kedua algoritma tersebut, maka penulis mendapatkan hasil sebagai berikut:

Tabel 1. *Confusion Matrix* Metode SVM menggunakan Haberman's Survival

Kinerja Klasifikasi	**Predicted Class**	
Actual Class	Predicted. Survived 5 years or longer	Predicted. Died within 5 years
Actual. Class Survived 5 years or longer	26 (True Positive)	2 (False Negative)
Actual. Class Died within 5 years	3 (False Positive)	0 (True Negative)

Berdasarkan tabel 1, maka dilanjutkan dengan menghitung nilai *Accuracy* pengklasifikasian dari model klasifikasi SVM menggunakan Dataset Haberman's Survival. Berikut hasil perhitungannya:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{26+0}{26+0+3+2} = \frac{26}{31} =$$

0.8387*100% = **83.87%**

Dengan demikian tingkat kedekatan antara nilai prediksi *class* dengan nilai aktual *class* atau jumlah prediksi *class* yang benar dari model klasifikasi SVM terhadap Dataset Haberman's Survival adalah sebesar 83.87%.

 ISSN 2089-5490

Tabel 2 *Confusion Matrix* Metode *Naïve Bayesian* menggunakan Haberman's Survival

Kinerja Klasifikasi	**Predicted Class**	
Actual Class	Predicted. Survived 5 years or longer	Predicted. Died within 5 years
Actual. Class Survived 5 years or longer	25 (True Positive)	3 (False Negative)
Actual. Class Died within 5 years	3 (False Positive)	0 (True Negative)

Berdasarkan tabel 1, maka dilanjutkan dengan menghitung nilai *Accuracy* pengklasifikasian dari model klasifikasi *Naïve Bayesian*menggunakan Dataset Haberman's Survival. Berikut hasil perhitungannya:

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ = $\frac{25+0}{25+0+3+3}$ = $\frac{25}{31}$ = 0.8065*100% = **80.65%**

Dengan demikian tingkat kedekatan antara nilai prediksi *class* dengan nilai aktual *class* atau jumlah prediksi *class* yang benar dari model klasifikasi *Naïve Bayesian* terhadap Dataset Haberman's Survival adalah sebesar 80.65%.

V. PERBANDINGAN *PERFORMANCE* METODE *DATASET*

Dari hasil pengujian diatas, maka dapat disimpulkan tingkat akurasi dari kedua algoritma tersebut yang dapat dilihat pada tabel berikut:

Tabel 3 Perbandingan Performance Metode

	SVM	Naïve Bayes
precision	89.66%	89.29%
class recall	92.86%	89.29%
Accuracy	83.87%	80.65%

Dari tabel diatas diketahui nilai Precision dari Support Vector Machine lebih besar dengan nilai 89.66% dari Naïve bayes yang menghasilkan 89.29% yang memiliki selisih 0.37 % dari kedua algoritma tersebut. Sedangkan Class Recall yang dihasilkan Naïve Bayes mencapai 89.29% dan Support Vector Machine 92.86% dan dengan perbedaan nilai akurasi mencapai 3.22% dari perbandingan performa kedua algoritma tersebut.

VI. HASIL

Pada penelitian yang dilakukan mengenai *Haberman's Survival dataset* dilakukan dengan menghasilkan prediksi dari metode *Support Vector Machine* dan *Naïve Bayes* dengan melakukan pencarian tingkat keakuratan tertinggi yang baik.

Dari hasil analisis bahwa *Support Vector Machine* mampu meningkatkan akurasi dari metode *Naïve Bayes*, dimana peningkatan rata-rata akurasi tertinggi terhadap *Support Vector Machine* diperoleh pada *Haberman's Survival dataset* yaitu sebesar 89.66% pada *class precision* dan sedangkan *Naïve Bayes* mencapai 89.29%. Pada *class recall Naïve Bayes* memperoleh nilai 89.29% dan *Support Vector Machine* 92.86%, dan akurasi data mencapai 80.65% pada *Naïve Bayes* dan 83.87% pada *Support Vector Machine*.

Keberhasilan dalam memprediksi menggunakan metode *Support Vector Machine* menggunakan *Haberman's Survival dataset.*

VII. DAFTAR PUSTAKA

[1] Albert, J. 2009. *Bayesian Computation with R*, Springer : New York.

[2] Bustami. 2013. *Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi*. TECHSI : Jurnal Penelitian Teknik Informatika. Vol. 3, No.2, Hal. 127-146.

[3] Bolstad, W.M. 2007. *Introduction to Bayesian Statistics*. John Wiley and Sons : New Jersey.

[4] Bramer, M. 2007. *Principles of Data Mining*. London: Springer.

[5] Gorunescu, F. 2011. *Data Mining: Concepts, Models and Techniques.* Berlin: Springer-Verlag.

[6] Han, J., Kamber, M. 2001. *"Data Mining Concepts and Techniques"*, Morgan Kaufman Pub., USA.

[7] Han, J. and Kamber, M. 2006. "Data Mining Concepts and Techniques Second Edition". Morgan Kauffman, San Francisco.

[8] Han, J., Kamber, M., & Pei, J. 2011. *Data Mining: Concepts and Techniques (3rd ed.).* San Francisco: Morgan Kaufmann Publishers Inc.

[9] Patil, T. R., Sherekar, M. S., 2013. *Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification*. International Journal of Computer Science and Applications, Vol. 6, No. 2, Hal 256-261.

 ISSN 2089-5490

[11] Pattekari, S. A., Parveen, A. 2012. *Prediction System for Heart Disease Using Naive Bayes*. International Journal of Advanced Computer and Mathematical Sciences, ISSN 2230-9624, Vol. 3, No 3, Hal 290-294.

[12] Ridwan, M., Suyono, H., Sarosa, M. 2013. *Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier,*. Jurnal EECCIS, Vol 1, No. 7, Hal. 59-64.

[13] Santosa, B. 2007. *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis.* Graha Ilmu. Yogyakarta.

[14] Zarlis, M., Sitompul, O.S., Sawaluddin, Effendi, S., Sihombing, P. & Nababan, E.B.

[15] 2015. Pedoman Penulisan Tesis. FasilkomTI. Universitas Sumatera Utara.

[16] Zhang, H., & Wang. Z. 2011. "A Normal Distributions-Based Over-Sampling Approach to Imbalanced Data Classification. *Advanced Data Mining and Applications – 7th International Conference* (pp. 83-96). Beijing, Springer.