Klasifikasi Pengidap Diabetes Pada Perempuan Menggunakan Penggabungan Metode Support Vector Machine dan K-Nearest Neighbour

Januar Adi Putra*, Afrizal Laksita Akbar**
Jurusan Teknik Informatika, Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember (ITS) – Surabaya, 60111, Indonesia
*januaradi.putra@gmail.com, **afrizal.la@gmail.com

ABSTRACT

Diabetes Mellitus is a metabolic disease with characteristics of hyperglycemia that occurs due to abnormalities in insulin secretion, insulin action or both. The detection of diabetes mellitus disease using the dataset Pima Indians had been done by various methods, one of which implementation methods is K-Neaarest Neighbor (KNN). One drawback of the KNN method is the determination of the optimal parameters k. Value of k that are too high will reduce the effect of noise on the classification, but makes the boundaries between each classification is becoming increasingly blurred, while the value of k that is too low will result in sample taking values for the less and lead to reduced accuracy. For this study proposes the use of Support Vector Machine (SVM) as the optimal solution of k determination. In this study, we will implement the hybrid SVM-KNN method to be used as a method of classification of people with diabetes using the dataset "Pima indian". Experiments done by varying the parameter values and the kernel used to see the value of the accuracy of the hybrid SVM-KNN method. Parameters that influence the value of C, tolerance, sigma, bias and the value of k on KNN. The highest average value of the accuracy obtained by using SVM-KNN is 92.00% and proved to be better than traditional SVM method average of the accuracy only 77.60% and KNN is 91%.

Keyword: Classification, SVM, KNN, Diabetes Millitus, Pima Indian

1. Introduction

Sejak 1991 setiap tahun, IDF (International Diabetes Federation) dan Organisasi Kesehatan Dunia (WHO) telah menetapkan tanggal 14 Nopember sebagai Hari Diabetes. Pada tahun 2007, hari diabetes ini resmi sebagai hari sedunia dalam agenda PBB. Ini menandakan kalau penyakit Diabetes Melitus (disingkat DM) tidak bisa dipandang sebelah mata. WHO memprediksi kenaikan jumlah penyandang DM di Indonesia dari 8,4 juta pada tahun 2000 menjadi sekitar 21,3 juta pada tahun 2030 [1]. Pada perkembangan di dunia kedokteran saat ini, peneliti dan praktisi memusatkan perhatiannya untuk mendeteksi DM dan mencegah atau menghambat berkembangnya komplikasi. Untuk mendeteksi seseorang terkena diabetes, ada beberapa tes lab yang harus dilakukan. US National Institute of Diabetes telah melakukan uji untuk penyakit diabetes sesuai dengan kriteria Organisasi Kesehatan Dunia yang dilakukan pada sejumlah perempuan yang berusia di atas 21 tahun, dari warisan Pima India dan tinggal di dekat Phoenix, Arizona, Amerika Serikat. Lebih dari 50% populasinya menderita DM. Dataset Pima meliputi delapan atribut pengukuran dari pasien yang DM positif dan pasien didiagnosis DM negatif.

Pendeteksian penyakit diabetes menggunakan dataset diabetes "Pima Indian" sudah pernah dilakukan dengan berbagai metode, salah satunya implementasi metode K-Neaarest Neighbour (KNN) yang digunakan oleh [2] pada data diabetes Pima Indian. Tujuan dari algoritma *k*-NN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan *training samples*, dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Akurasi tertinggi yang didapatkan pada penelitian ini mencapai 91%. Salah satu kelemahan dari metode KNN adalah dalam penentuan parameter k yang optimal, nilai k yang terlalu tinggi akan mengurangi efek noise pada klasifikasi, tetapi membuat batasan antara setiap klasifikasi menjadi semakin kabur, sedangkan nilai k yang terlalu rendah akan mengakibatkan sample pengambilan nilai pembanding semakin sedikit dan menyebabkan keakurasian berkurang[3]. Untuk itu penelitian ini menggunakan Support Vector Machine (SVM) sebagai solusi penentuan k. Berdasar permasalahan yang telah dipaparkan, kami akan menerapkan metode hybrid SVM-KNN untuk digunakan sebagai metode klasifikasi pengidap diabetes menggunakan data pima indian, nantinya hasil akurasi dari metode hybrid ini akan dibandingkan dengan hasil akurasi metode SVM dan KNN tradisionil.

2. Research Method

2.1 Diabetes Melitus (DM)

Diabetes Melitus merupakan penyakit metabolik dengan karakteristik hiperglikemia yang terjadi karena kelainan sekresi insulin, kerja insulin atau keduanya. jika telah terkena kronik diabetes, maka akan terjadi kerusakan jangka panjang, disfungsi atau kegagalan beberapa organ tubuh terutama mata ginjal, mata, saraf, jantung dan pembuluh darah [4]. Orang yang sehat memiliki beberapa hormon insulin bertugas mengatur kadar glukosa darah. Insulin diproduksi oleh pankreas, organ kecil dekat perut yang juga mengeluarkan enzim penting yang membantu dalam proses pencernaan makanan. Insulin mengatur glukosa untuk bergerak dari darah ke dalam hati, otot dan sel lemak dimana ini digunakan untuk bahan bakar.

2.2 Data Diabetes Pima Indian

Dataset yang digunakan pada penelitian ini diambil dari repositori database UCI Pima Indian (http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes). Dataset Pima Indian terdiri dari 768 data diagnosis DM. Data ini memiliki 8 atribut dengan target output positif diabetes (ditunjukkan dengan output 1) dan negatif diabetes (ditunjukkan dengan output 0). Daftar atribut data diabetes ditunjukkan pada Tabel 1.

Tabel 1. Daftar atribut pima indian

Atribut	Singkatan	Deskripsi	Satuan	Type Data
Pregnant	Hamil	Banyaknya kehamilan	-	Continous
Plasma Glucose Concentration	OGTT	Kadar glukosa 2 jam setelah makan	mg/dl	Continous
Diastolic Blood Pressure	Diastolik	Tekanan darah	mmhg	Continous
Triceps Skin Fold Thickness	TSFT	Ketebalan kulit	mm	Continous
2-Hour Serum Insulin	INS	Insulin	mu U/ml	Continous
Body Mass Index	IMB	Berat tubuh	kg/m^2	Continous
Diabetes Pedigree Function	DPF	Riwayat diabetes dalam kelarga	-	Continous
Age	Usia	Umur pasien	Tahun	Continous

2.3 K-Nearest Neighbour (KNN)

Algoritma k-NN adalah suatu metode yang menggunakan algoritma *supervised* [5], [6],[7],[8],[9]. Perbedaan antara *supervised learning* dengan *unsupervised learning* adalah pada *supervised learning* bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang sudah ada dengan data yang baru. Sedangkan pada *unsupervised learning*, data belum memiliki pola apapun, dan tujuan *unsupervised learning* untuk menemukan pola dalam sebuah data[5], [6], [7], [8].

Tujuan dari algoritma *k*-NN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan *training samples* [6],[7]. Dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada *k*-NN. Pada proses pengklasifikasian, algoritma ini tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori. Algoritma *k*-NN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari sampel uji yang baru [5], [6], [7]. Jarak yang digunakan adalah jarak *Euclidean Distance*. Jarak *Euclidean* adalah jarak yang paling umum digunakan pada data numerik [10]. *Euclidean distance* didefinisikan sebagai berikut [7] :

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \quad (1)$$

Keterangan :

$d(x_i, x_j)$: Jarak Euclidean (*Euclidean Distance*)

(x_i) : record ke-i

(x_j) : record ke-j

(a_r) : data ke-r

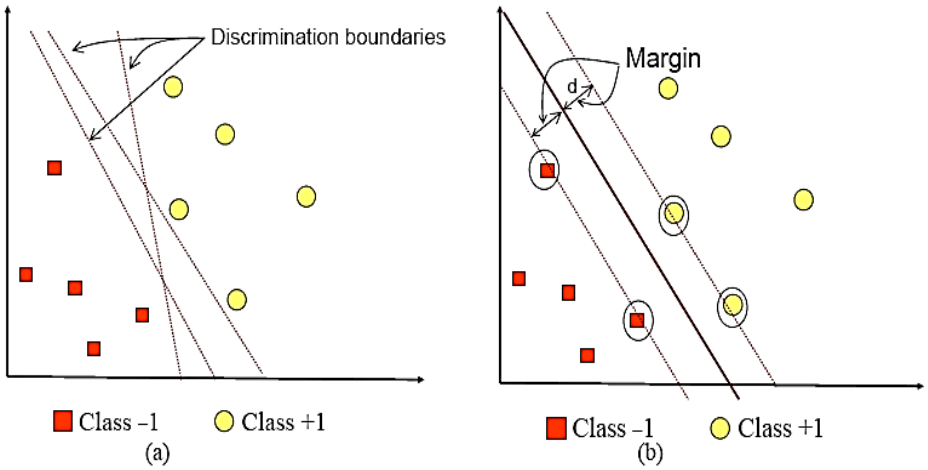
i, j : 1,2,3,....n

Algoritma *k*-NN adalah algoritma yang menentukan nilai jarak pada pengujian *data testing* dengan *data training* berdasarkan nilai terkecil dari nilai ketetanggaan terdekat [10] didefinisikan sebagi berikut:

$$D_{nn}(C_1, C_2) = \min_{1 \le i \le r, 1 \le j \le s} d(y_i, Z_j) \quad (2)$$

2.4 Support Vector Machine (SVM)

Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah class pada input space. Gambar 1a memperlihatkan beberapa pattern yang merupakan anggota dari dua buah class : positif (dinotasikan dengan +1) dan negatif (dinotasikan dengan –1). Pattern yang tergabung pada class negatif disimbolkan dengan kotak, sedangkan pattern pada class positif, disimbolkan dengan lingkaran. Proses pembelajaran dalam problem klasifikasi diterjemahkan sebagai upaya menemukan garis (hyperplane) yang memisahkan antara kedua kelompok tersebut. Berbagai alternatif garis pemisah (discrimination boundaries) ditunjukkan pada Gambar 1a.



Gambar 1. SVM berusaha menemukan hyperplane terbaik

Hyperplane pemisah terbaik antara kedua class dapat ditemukan dengan mengukur margin hyperplane tsb. dan mencari titik maksimalnya. Margin adalah jarak antara hyperplane tersebut dengan data terdekat dari masing-masing class. Subset data training set yang paling dekat ini disebut sebagai support vector. Garis solid pada Gambar 1b menunjukkan hyperplane yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua class, sedangkan titik kotak dan lingkaran yang berada dalam lingkaran hitam adalah support vector. Upaya mencari lokasi hyperplane optimal ini merupakan inti dari proses pembelajaran pada SVM. Data yang tersedia dinotasikan sebagai $\vec{x_i} \in \Re^d$ sedangkan label masing-masing dinotasikan $y_i \in \{-1,+1\}$ untuk i = 1, 2,…,l , yang mana l adalah banyaknya data. Diasumsikan kedua class –1 dan +1 dapat terpisah secara sempurna oleh hyperplane berdimensi d , yang didefinisikan:

$$\vec{w}.\vec{x} + b = 0 \quad (3)$$

Sebuah *pattern* $\vec{x_i}$ yang termasuk class –1 (sampel negatif) dapat dirumuskan sebagai *pattern* yang memenuhi pertidaksamaan:

$$\vec{w}.\vec{x} + b \le -1 \quad (4)$$

Sedangkan pattern $\vec{x_i}$ yang termasuk class +1 (sampel positif):

$$\vec{w}.\vec{x} + b \ge +1 \quad (5)$$

Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara hyperplane dan titik terdekatnya, yaitu $1/\left\|\vec{w}\right\|$. Hal ini dapat dirumuskan sebagai Quadratic Programming (QP) problem, yaitu mencari titik minimal persamaan (6), dengan memperhatikan constraint persamaan (7).

$$\min_{\vec{w}} \tau(w) = \frac{1}{2}\left\|\vec{w}\right\|^2 \tag{6}$$

$$y_i(\vec{x_i}.\vec{w}+b) = \frac{1}{2}\left\|\vec{w}\right\|^2 \tag{7}$$

Problem ini dapat dipecahkan dengan berbagai teknik komputasi, di antaranya Lagrange Multiplier sebagaimana ditunjukkan pada persamaan (8).

$$L(\vec{w},b,\alpha) = \frac{1}{2}\left\|\vec{w}\right\|^2 - \sum_{i=1}^{l}\alpha_i(y_i((\vec{x}.\vec{w}+b)-1)) \qquad (i=1,2,...l) \tag{8}$$

α_i adalah *Lagrange multipliers*, yang bernilai nol atau positif $(\alpha_i \geq 0)$. Nilai optimal dari persamaan (8) dapat dihitung dengan meminimalkan L terhadap w dan b, dan memaksimalkan L terhadap α_i. Dengan memperhatikan sifat bahwa pada titik optimal gradient L =0, persamaan (8) dapat dimodifikasi sebagai maksimalisasi problem yang hanya mengandung α_i saja, sebagaimana persamaan (9).

Maximize:

$$\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j \vec{x_i}\vec{x_j} \tag{9}$$

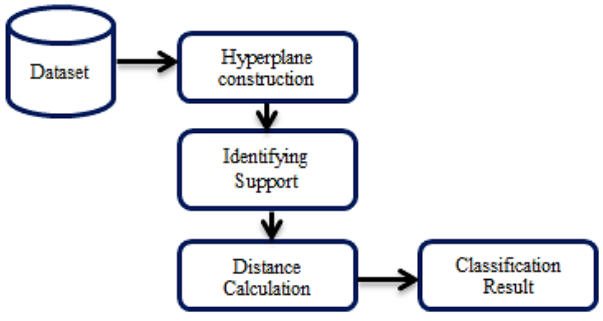
Subject to:

$$\alpha_i \geq 0(i=1,2,..l)\sum_{i=1}^{l}\alpha_i y_i = 0 \tag{10}$$

Dari hasil dari perhitungan ini diperoleh α_i yang kebanyakan bernilai positif. Data yang berkorelasi dengan α_i yang positif inilah yang disebut sebagai ***support vector***.

2.5 Hybrid SVM-KNN

Penelitian ini menggunakan metode hybrid SVM-KNN pada data diabetes pima indian dan alur metode usulan yang digunakan terlihat pada Gambar 2. Metode SVM digunakan untuk mengidentifikasi *support vector* yang membedakan antara kelas satu dengan yang lainnya. Setelah didapatkan maka *support vector* tersebut digunakan sebagai tetangga terdekat (k) dari k-nearest neighbour dan dilakukan perhitungan euclidean distance, dimana kelas dari tetangga yang terdekat atau dengan distance terkecil merupakan kelas dari data uji.



Gambar 2. Diagram alur pada klasifikasi dokumen teks

3. Result and Analysis

Telah dilakukan ujicoba untuk mendapatkan nilai akurasi dari metode usulan. Pada ujicoba yang telah dilakukan dataset yang digunakan adalah dataset pima indian. Database diabetes Pima India, disumbangkan oleh Vincent Sigillito. Data Diabetes India Pima adalah kumpulan laporan diagnostik medis dari 768 contoh-contoh dari populasi yang tinggal di dekat Phoenix, Arizona, Amerika Serikat. Jenis pengujian yang dilakukan adalah pengujian *trainingtest* dimana semua data training dijadikan sebagai data uji. Hasil dari ujicoba dapat dilihat pada Tabel 2,3,4,5,6, dan 7. Ujicoba dilakukan dengan mengubah-ubah nilai parameter dan kernel yang digunakan untuk melihat nilai akurasi dari metode hybrid SVM-KNN. Dari ujicoba yang dilakukan dapat dilihat bahwa akurasi tertinggi didapat pada ujicoba ke-IV dengan nilai akurasi 92.00% sedangkan nilai akurasi terendah terjadi pada ujicoba ke-V dengan nilai akurasi 86.55%. akurasi tertinggi didapatkan pada nilai parameter C=7, toleransi=0.7, sigma=7, bias=7, dan Pangkat=7. Perbandingan akurasi juga dilakukan untuk mengatahui apakah metode usulan lebih baik dari SVM dan KNN tradisionil, hasil perbandingan akurasi dapat dilihat pada Tabel 8. Dimana terlihat metode usulan lebih baik dari metode SVM dan KNN tradisionil.

Tabel 2. Ujicoba I

Kernel	C	Toleransi	Sigma	Bias	Pangkat	Akurasi %
Linear	1	0.1	1	1	1	100.00
Polynomial	1	0.1	1	1	1	100.00
Gaussian	1	0.1	1	1	1	100.00
Tanh	1	0.1	1	1	1	63.90
Rata-rata						**90.98**

Tabel 3. Ujicoba II

Kernel	C	Toleransi	Sigma	Bias	Pangkat	Akurasi %
Linear	3	0.3	3	3	3	100.00
Polynomial	3	0.3	3	3	3	89.70
Gaussian	3	0.3	3	3	3	100.00
Tanh	3	0.3	3	3	3	64.95
Rata-rata						**88.67**

Tabel 4. Ujicoba III

Kernel	C	Toleransi	Sigma	Bias	Pangkat	Akurasi %
Linear	5	0.5	5	5	5	100.00
Polynomial	5	0.5	5	5	5	90.00
Gaussian	5	0.5	5	5	5	100.00
Tanh	5	0.5	5	5	5	65.74
Rata-rata						**88.94**

Tabel 5. Ujicoba IV

Kernel	C	Toleransi	Sigma	Bias	Pangkat	Akurasi %
Linear	7	0.7	7	7	7	100.00
Polynomial	7	0.7	7	7	7	100.00
Gaussian	7	0.7	7	7	7	100.00
Tanh	7	0.7	7	7	7	68.00
Rata-rata						**92.00**

Tabel 6. Ujicoba V

Kernel	C	Toleransi	Sigma	Bias	Pangkat	Akurasi %
Linear	9	0.9	9	9	9	100.00
Polynomial	9	0.9	9	9	9	85.10
Gaussian	9	0.9	9	9	9	100.00
Tanh	9	0.9	9	9	9	61.13
Rata-rata						**86.55**

Tabel 7. Ujicoba VI

Kernel	C	Toleransi	Sigma	Bias	Pangkat	Akurasi %
Linear	1	0.001	5	3	2	100.00
Polynomial	1	0.001	5	3	2	90.80
Gaussian	1	0.001	5	3	2	100.00
Tanh	1	0.001	5	3	2	65.35
Rata-rata						**89.03**

Tabel 8. Perbandingan dengan metode lain

Metode	Akurasi %
Usulan	**92.00**
SVM [11]	77.60
KNN [2]	91.00

4. Conclusion

Dari ujicoba yang telah dilakukan metode usulan terbukti mampu mengoptimallkan pemilihan parameter k dari metode KNN, sehingga nilai akurasi yang didapat dapat lebih tinggi dari metode KNN dan SVM tradisionil. Faktor yang mempengaruhi metode usulan sama halnya dengan faktor yang mempengaruhi metode SVM yakni parameter C, toleransi, sigma, bias dan pangkat, kernel yang dipilih pun dapat mempengaruhi tingkat ketepatan dalam mengklasifikasi.

Acknowledgements (10 PT)

Penulis ucapkan terimakasih kepada segenap dosen pengajar di Teknik Informatika ITS terutama Ibu Dr. Chastine Fatichah, M.Kom. selaku pembimbing dalam penelitian yang telah dilakukan hingga tersusun jurnal tentang *machine learning* ini.

References

[1] World Health Organization Department of Noncommunicable Disease Surveillance. (1999). *Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications*.Geneva: Department of Noncommunicable Disease Surveillance

[2] Rahman A.R, Nurhayati, *Implementation of Naive Bayes and K-Nearest Neighbor Algorithm for Diagnosis of Diabetes Mellitus*, Applied Computational Science; ISBN: 978-960-474-368-1.

[3] Vinoth.R,et al. *A Hybrid Text Classification Approach Using KNN and SVM*. International Journal of Advance Foundation and Research in Computer (IJAFRC).2014

[4] Regina. 2012. *Penyakit Diabetes Melitus*. http://diabetesmelitus.org, diakses tanggal 23 Desember 2015

[5] Wu X, Kumar V. *The Top Ten Algorithms in Data Mining*. New York:CRC Press;2009

[6] Larose D. *Discovering Knowledge in Data*. USA:John Wiley's and Son ;2005

[7] Han J and Kamber M. *Data Mining:Concept and Techniques*. New York:Morgan Kaufmann Publisher;2006

[8] Mitsa T. *Temporal Data Mining*. New York :CRC Press;2010.

[9] Nugroho A. *k-Nearest Neighbor (k-NN)*. 2010 [Updated 2011 Mei 2; diakses 2015 Desember 24]. Available from:Http://asnugroho.Wordpress.com/ 2007/01/26/k-nearest-neighbor-classifier/

[10] Goujon G, Chaoqun, Jianhong W. *Data Clusterin :Theory, Algorithms, and Applications*. Virginia: ASA;2007.

[11] Parashar.A, Barse.K, Rawat.K. *A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA-Support Vector Machine and Feed Forward Neural Network. IJARCSSE.Volume 4, Issue 11, November 2014*