Jurnal Algoritma, Logika dan Komputasi *Vol.VII (No. 01) : 652 – 660 Th. 2024*
Versi Online: https://journal.ubm.ac.id/index.php/alu *p-ISSN: 2620-620X*
DOI: http://dx.doi.org/10.30813/j-alu.v2i2.6029 *e-ISSN: 2621-9840*

MODEL KLASIFIKASI HIBRIDA BARU DARI JARINGAN SYARAF TIRUAN DAN MODEL REGRESI LINIER BERGANDA

NEW HYBRID CLASSIFICATION MODEL OF ARTIFICIAL NEURAL NETWORKS AND MULTIPLE LINEAR REGRESSION MODELS

Andre Valerian, miukuchan1@gmail.com[1)] & Honni, honni2482@gmail.com[2)*]
[1)]Program Studi Informatika, Fakultas Teknik dan Desain, Universitas Bunda Mulia
[2)]Program Studi Sistem Informasi, Fakultas Teknik dan Desain, Universitas Bunda Mulia

Diterima 26 Maret 2024 / Disetujui 31 Juli 2024

ABSTRACT

This paper examines a more accurate and broader classification model and has significant implications in these fields. Combining multiple models or using hybrid models has become common practice to overcome the shortcomings of a single model and can be a more effective way to improve its predictive performance, especially when the models are in very different combinations. In this paper, a new hybridization of artificial neural networks (ANN) is proposed using multiple linear regression models to produce more accurate models than traditional artificial neural networks for solving classification problems. Empirical results show that the proposed hybrid model shows to effectively improve classification accuracy compared to traditional artificial neural networks and also several other classification models such as linear discriminant analysis, quadratic discriminant analysis, and vector machine using benchmarks and real-world application datasets. These datasets vary in number of classes and data sources. Therefore, it can be applied as a suitable alternative approach to solve classification problems, especially when higher forecasting accuracy is required.
***Keywords**: Artificial Neural Network, Classification model, Linear regression model.*

ABSTRAK

Paper ini mengkaji model klasifikasi yang lebih akurat dan lebih luas serta memiliki implikasi yang signifikan dalam bidang-bidang ini. Menggabungkan beberapa model atau menggunakan model hibrida telah menjadi praktik umum untuk mengatasi kekurangan model tunggal dan dapat menjadi suatu cara yang lebih efektif untuk meningkatkan kinerja prediktif tersebut, terutama ketika model dalam kombinasi yang sangat berbeda. Dalam tulisan ini, hibridisasi baru dari jaringan saraf tiruan (JST) diusulkan menggunakan model regresi linier berganda untuk menghasilkan model yang lebih akurat daripada jaringan saraf tiruan tradisional untuk memecahkan masalah klasifikasi. Hasil empiris menunjukkan bahwa model hibrida yang diusulkan menunjukkan secara efektif meningkatkan akurasi klasifikasi dibandingkan dengan jaringan saraf tiruan tradisional dan juga beberapa model klasifikasi lain seperti analisis diskriminan linier, analisis diskriminan kuadrat, dan vector machine menggunakan patokan dan kumpulan data aplikasi dunia nyata. Set data ini bervariasi dalam jumlah kelas dan sumber data. Oleh karena itu, dapat diterapkan sebagai pendekatan alternatif yang tepat untuk memecahkan masalah klasifikasi, khususnya ketika akurasi peramalan yang lebih tinggi diperlukan.
Kata Kunci : Jaringan Saraf Tiruan, Model klasifikasi, Model regresi linier.

PENDAHULUAN

Klasifikasi adalah bidang penting dari penelitian yang berkaitan dengan menugaskan suatu objek ke salah satu dari satu set kelas, berdasarkan atribut dari objek itu. Kinerja proses klasifikasi tergantung pada seberapa baik fungsi diskriminan untuk masalah khusus. Diskriminan dikembangkan untuk meminimalkan tingkat kesalahan klasifikasi, dari beberapa sampel yang diberikan dari pasangan vektor input dan output, yang disebut sebagai kumpulan data pelatihan. Fungsi diskriminan ini kemudian digunakan untuk mengklasifikasikan pengamatan baru ke dalam kelompok yang didefinisikan sebelumnya dan untuk menguji akurasi klasifikasi. Masalah klasifikasi telah diperiksa di berbagai bidang seperti bisnis, obat-obatan, biologi, pengenalan citra, dll. Dan penggunaan model-model ini telah menjadi sangat diperlukan di bidang-bidang yang telah disebutkan di atas, terutama

*Korespondensi Penulis:
E-mail: **honni2482@gmail.com**

dalam bisnis dan keuangan. Pendekatan klasifikasi umumnya dikategorikan dalam dua kategori utama, pendekatan linear dan nonlinier.

Pendekatan klasifikasi linear mempartisi ruang input menjadi kumpulan daerah yang terpisah, dipisahkan oleh batas keputusan linier. Contoh terkenal dari teknik klasifikasi linier yang telah banyak digunakan dalam klasifikasi termasuk yang dengan regresi linier berganda, analisis diskriminan linier, regresi logistik, memisahkan bidang hiper, dll.

Teknik klasifikasi ini bekerja dengan baik ketika kelas terpisah secara linier . Namun, dalam banyak masalah dunia nyata data mungkin tidak dapat dipisahkan secara linier dan juga data sangat berdekatan dan oleh karena itu diperlukan batas keputusan yang sangat nonlinier untuk memisahkan data[1]. Beberapa kelas teknik klasifikasi nonlinier telah diusulkan dalam literatur untuk mengatasi keterbatasan linear dari teknik klasifikasi linier. Teknik-teknik ini termasuk teknik klasik seperti analisis diskriminan kuadrat, dll. Dan pendekatan jaringan saraf tiruan seperti pohon neural, multilayer perceptrons, jaringan saraf probabilistik, vector machine, dll. Beberapa fitur yang membedakan jaringan saraf tiruan yang membuat menarik adalah Pertama, dibandingkan dengan teknik berbasis model tradisional, jaringan syaraf tiruan adalah metode self-adaptif data-driven karena hanya ada sedikit asumsi untuk model masalah yang diteliti. Kedua, jaringan saraf tiruan dapat digeneralisasikan. Setelah mempelajari data yang disajikan kepada mereka (sampel), jaringan syaraf tiruan sering dapat atau dengan tepat menyimpulkan bagian populasi yang tidak terlihat bahkan jika data sampel mengandung informasi yang tidak akurat. Ketiga, jaringan saraf tiruan adalah aproksimator fungsional universal. Telah ditunjukkan bahwa suatu jaringan dapat memperkirakan fungsi kontinu untuk keakuratan yang diinginkan. Akhirnya, jaringan saraf tiruan bersifat nonlinear [2][3].

Definisi Jaringan Saraf Tiruan

- *Hecht-Nielsend (1988),* "Suatu neural network (NN), adalah suatu struktur pemroses informasi yang terdistribusi dan bekerja secara paralel, yang terdiri atas elemen pemroses (yang memiliki memori lokal dan beroperasi dengan informasi lokal) yang diinterkoneksi bersama dengan alur sinyal searah yang disebut koneksi. Setiap elemen pemroses memiliki koneksi keluaran tunggal yang bercabang (fan out) ke sejumlah koneksi kolateral yang diinginkan (setiap koneksi membawa sinyal yang sama dari keluaran elemen pemroses tersebut). Keluaran dari elemen pemroses tersebut dapat merupakan sebarang jenis persamaan matematis yang diinginkan. Seluruh proses yang berlangsung pada setiap elemen pemroses harus benar-benar dilakukan secara lokal, yaitu keluaran hanya bergantung pada nilai masukan pada saat itu yang diperoleh melalui koneksi dan nilai yang tersimpan dalam memori lokal".[4]
- *Haykin, S. (1994),* Sebuah jaringan saraf adalah sebuah prosesor yang terdistribusi paralel dan mempuyai kecenderungan untuk menyimpan pengetahuan yang didapatkannya dari pengalaman dan membuatnya tetap tersedia untuk digunakan. Hal ini menyerupai kerja otak dalam dua hal yaitu: 1. Pengetahuan diperoleh oleh jaringan melalui suatu proses belajar. 2. Kekuatan hubungan antar sel saraf yang dikenal dengan bobot sinapsis digunakan untuk menyimpan pengetahuan.[5]
- *Zurada, J.M. (1992),* Sistem saraf tiruan atau jaringan saraf tiruan adalah sistem selular fisik yang dapat memperoleh, menyimpan dan menggunakan pengetahuan yang didapatkan dari pengalaman.[6]
- DARPA Neural Network Study (1988), Sebuah jaringan syaraf adalah sebuah sistem yang dibentuk dari sejumlah elemen pemroses sederhana yang bekerja secara paralel dimana fungsinya ditentukan oleh stuktur jaringan, kekuatan hubungan, dan pegolahan dilakukan pada komputasi elemen atau nodes[7]
- JJ Siang, sistem pemrosesan informasi yang memiliki karakteristik mirip dengan jaringan syaraf manusia.[8]

Asumsi Jaringan Saraf Tiruan

Jaringan syaraf tiruan dibentuk sebagai generalisasi model matematika dari jaringan syaraf manusia, dengan asumsi JST:[9]

- Pemrosesan terjadi pada banyak elemen yang sederhana

- Sinyal dikirim diantara neuron2 melalui sinapsis
- Sinapsis memiliki bobot yang akan memperkuat atau memperlemah sinyal.
- Output ditentukan menggunakan fungsi aktivasi yang dikenakan pada jumlah input yang diterima
- Output dibandingan dengan suatu tracehold.

Pengertian Analisis Regresi.

Analisis Regresi adalah analisis yang mengukur pengaruh variabel bebas terhadap variabel terikat. Pengukuran pengaruh ini melibatkan satu variabel bebas (X) dan variabel terikat (Y), yang dinamakan analisis regresi linier sederhana dengan rumus Y= a+bX. Nilai "a" adalah konstanta dan nilai "b" adalah koefisien regresi untuk variabel X.[10]
Harga 'a' dapat dicari dengan rumus (1)

$$a = \frac{\sum Y(\sum X^2)\ \sum X.\sum Y}{n\sum X^2 - (\sum X)^2} \quad (1)$$

Harga 'b' dapat dicari dengan rumus (2)

$$b = \frac{n\sum XY - \sum X.\sum XY}{n\sum X^2 - (\sum X)^2} \quad (2)$$

Koefisien regresi 'b' adalah kontribusi besarnya perubahan nilai **variabel bebas**, semakin besar nilai koefisien regresi maka kontribusi perubahan semakin besar, demikian pula sebaliknya akan semakin lecil. Kontribusi perubahan variabel bebas (X) juga ditentukan oleh koefisien regresi positif atau negatif. [10]

Pengukuran Analisis Regresi

Pengukuran pengaruh variabel yang melibatkan lebih dari satu variabel bebas (X1,X2,X3,…,Xn), digunakan analisis regresi linier berganda, disebut linier karena setiap estimasi atas nilai diharapkan memgalami peningkatan atau penurunan mengikuti garis lurus. Rumus (3) merupakan estimasi regresi linier berganda :[11]

$$Y=a+b_1X_1+b_2X_2+b_3X_3+\ldots+b_nX_n \quad (3)$$

Keterangan :
Y : variabel terikat *(dependent)*
X (1,2,3,…) : variabel bebas *(independent)*
a : nilai konstanta
b (1,2,3,…) : nilai koefisien regresi
Penggunaan nilai konstanta secara statistik dilakukan jika satuan-satuan variabel X *(independent)* dan variabel Y *(dependent)* tidak sama. Sedangkan, bila variabel X *(independent)* dan variabel Y *(dependent)*, baik linier sederhana maupun berganda, memiliki satuan yang sama maka nilai konstanta diabaikan dengan asumsi perubahan variabel Y *(dependent)* akan proposional dengan nilai perubahan variabel X *(independent)*.

Dalam menentukan nilai 'a' dan 'b1′,'b2′,'b3',.., digunakan persamaan regresi linier berganda:

1. SY = an+b1SX1+b2SX2+b3SX3+…
2. SX1Y = aSX1+b1SX1²+b2SX1X2+…
3. SX2Y = aSX2+b2SX1X2+b2SX21²+…

dan seterusnya. Untuk menghitung nilai 'a','b1′,'b2′,'b3′,… pada persamaan regresi linier berganda dapat dirumuskan =nx-1 di mana nx = banyaknya variabel bebas (X).

Definisi K-Nearest Neighbors (k-NN)

K-nearest neighbors atau knn adalah algoritma yang berfungsi untuk melakukan klasifikasi suatu data berdasarkan data pembelajaran (*train data sets*), yang diambil dari k tetangga terdekatnya (*nearest neighbors*). Dengan k merupakan banyaknya tetangga terdekat.[12]

Cara Kerja Algoritma K-Nearest Neighbors (KNN)

K-nerest neaighbors melakukan klasifikasi dengan proyeksi data pembelajaran pada ruang berdimensi banyak. Ruang ini dibagi menjadi bagian-bagian yang merepresentasikan kriteria data pembelajaran. Setiap data pembelajaran direpresentasikan menjadi titik-titik ***c*** pada ruang dimensi banyak.[13]

Klasifikasi Terdekat (Nearest Neighbor Classification)

Data baru yang diklasifikasi selanjutnya diproyeksikan pada ruang dimensi banyak yang telah memuat titik-titik c data pembelajaran. Proses klasifasikasi dilakukan dengan mencari titik ***c*** terdekat dari ***c-baru*** (*nearest neighbor*). Teknik pencarian tetangga terdekat yang umum dilakukan dengan menggunakan formula jarak euclidean. Berikut beberapa formula yang digunakan dalam algoritma knn.[14]

- Euclidean Distance

Jarak Euclidean adalah formula untuk mencari jarak antara 2 titik dalam ruang dua dimensi[15] dapat dilihat pada Rumus (4)

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (4)$$

- Hamming Distance

Jarak Hamming adalah cara mencari jarak antar 2 titik yang dihitung dengan panjang vektor biner yang dibentuk oleh dua titik tersebut dalam block kode biner.[16]

- Manhattan Distance

Manhattan Distance atau Taxicab Geometri adalah formula untuk mencari jarak ***d*** antar 2 vektor ***p,q*** pada ruang dimensi ***n[17]***.

- Minkowski Distance

Minkowski distance adalah formula pengukuran antar 2 titik pada ruang vektor normal yang merupakan hibridisasi yang mengeneralisasi euclidean distance dan mahattan distance.

Teknik pencarian tetangga terdekat disesuaikan dengan dimensi data, proyeksi, dan kemudahan implementasi oleh pengguna.[18]

HASIL DAN PEMBAHASAN

Model Perumusan hybrid

Terlepas dari banyaknya model klasifikasi yang tersedia, akurasi merupakan hal mendasar bagi banyak proses keputusan, dan karenanya, tidak ada yang meneliti cara-cara untuk meningkatkan keefektifan model-model klasifikasi yang telah diberikan. Banyak peneliti telah menggabungkan prediksi dari beberapa pengklasifikasi untuk menghasilkan pengklasifikasi yang lebih baik, yang telah dilaporkan untuk meningkatkan kinerja[17]. Efektivitas hibrida bergantung pada sejauh mana pengklasifikasiannya membuat kesalahan yang berbeda, atau tidak independen. Kesalahan berasal dari empat aspek, yaitu metode sampling data yang berbeda, pengaturan parameter yang berbeda, pengklasifikasi berbeda, dan strategi kombinasi yang berbeda[6]. Dengan menggunakan prediksi gabungan dari beberapa pengklasifikasi, kinerja yang lebih baik daripada penggolong individu dicari.

Breiman mengacu pada beberapa ahli pengklasifikasi yang telah menunjukkan potensi untuk mengurangi kesalahan generalisasi dari model classifier dari 5% hingga 70%. Dengan kata lain, beberapa pengklasifikasi dapat memberikan hasil klasifikasi yang lebih akurat daripada pengklasifikasi tunggal.[13]

$$y_t = w_0 + \sum_{j=1}^{q} w_j \cdot g\left(w_{0j} + \sum_{i=1}^{n} w_{ij} \cdot x_{t,i} + w_{n+1j} \cdot x_{t,n+1}\right) + \varepsilon_t$$
$$= \sum_{j=0}^{q} w_j \cdot g\left(w_{0j} + \sum_{i=1}^{n+1} w_{ij} \cdot x_{t,i}\right) + \varepsilon_t,$$

Model hibrida baru dari jaringan saraf tiruan diusulkan untuk menghasilkan hasil yang lebih akurat menggunakan model regresi linier berganda. Tujuan utama dari model yang diusulkan adalah untuk menggunakan keuntungan unik dari model regresi linier berganda dalam pemodelan linier untuk mengatasi keterbatasan pemodelan linear dari jaringan saraf tiruan tradisional. Oleh karena itu, pada fase pertama dari model yang diusulkan, model regresi linier berganda digunakan untuk memperbesar komponen linier dalam atribut untuk penggunaan yang lebih baik oleh jaringan saraf pada fase kedua. Kemudian komponen linier diperbesar dirangkum dalam atribut baru sebagai L (atribut n + 1th). Tujuan utama menggunakan model regresi linier berganda adalah untuk mengevaluasi hubungan antara atribut sebagai variabel independen atau variabel prediktor dan kelas sebagai variabel dependen. Ini dilakukan dengan memasang garis lurus ke sejumlah observasi. Secara khusus, garis diproduksi sehingga penyimpangan kuadrat dari titik-titik yang diamati dari garis yang diminimalkan. Dengan demikian prosedur ini umumnya disebut sebagai estimasi kuadrat terkecil[15]. Secara matematis, jika nilai kelas adalah linearitas bergantung pada nilai atributnya, maka model regresi berganda adalah sebagai berikut:

$$L = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = \sum_{i=0}^{n} \alpha_i x_i,$$

di mana xi (i = 0, 1, 2, ..., n) adalah atribut dan ai (i = 0, 1, 2, ..., n) adalah koefisien tidak diketahui yang diperkirakan dengan metode kuadrat terkecil. Kemudian, pada fase kedua dari model yang diusulkan.

Kemudian, pada tahap kedua, jaringan saraf digunakan untuk bersama- sama memodelkan struktur linier dan nonlinier dan mengklasifikasikan menggunakan atribut asli dan atribut linier yang dihasilkan oleh regresi linier berganda sebagai berikut:

di mana, $g\left(w_{0,0} + \sum_{i=1}^{n+1} w_{i,0} . x_{t,i}\right) = 1, w_{ij} (i = 0, 1, 2, \ldots, n + 1, j = 0, 1, 2, \ldots, q)$ dan $w_j (j = 0, 1, 2, \ldots, q)$ adalah bobot koneksi, n + 1 adalah nomor dari semua atribut (node input), dan q adalah jumlah node tersembunyi. Meskipun, dalam makalah ini, model yang diusulkan hanya digunakan untuk membangun model hibrida dengan multilayer perceptrons (MLP) untuk tujuan klasifikasi, metodologi ini secara umum dapat diterapkan pada berbagai jaringan syaraf tiruan seperti support vector machine (SVM), umum jaringan saraf regresi, jaringan saraf probabilistik, dll untuk pemodelan, perkiraan santai, dan tujuan klasifikasi.

Model klasifikasi dua kelas

Masalah klasifikasi berbeda karena outputnya berbeda. Namun, klasifikasi juga dapat dilihat sebagai proses menggambar partisi antara kelas. Model yang diusulkan dapat digunakan untuk memperkirakan fungsi yang mengidentifikasi partisi ini. Model yang kami usulkan tidak mengasumsikan bentuk partisi, tidak seperti analisis diskriminan linear dan kuadratik. Berbeda dengan metode tetangga K-terdekat, model yang diusulkan tidak memerlukan penyimpanan data

pelatihan. Setelah model telah dilatih, ia melakukan jauh lebih cepat daripada KNN karena tidak perlu iterate melalui sampel pelatihan individu. Model yang diusulkan tidak memerlukan eksperimen dan pemilihan akhir fungsi kernel dan parameter penalti seperti yang dibutuhkan oleh mesin vektor pendukung. Model yang kami usulkan hanya bergantung pada proses pelatihan untuk mengidentifikasi model penggolong akhir.

Untuk menerapkan model yang diusulkan untuk klasifikasi, modifikasi tertentu pada model perlu dibuat. Seperti halnya model klasifikasi lain (dengan pengecualian KNN), output dari model yang diusulkan bersifat kontinu, sementara klasifikasi membutuhkan hasil yang berbeda. Mirip dengan model lain, keluaran kontinu dari model yang diusulkan diubah ke kelas diskrit dengan menetapkan sampel ke kelas yang outputnya paling dekat. Setiap kelas diberi nilai numerik. Perbedaan antara output dan masing-masing nilai numerik kemudian dihitung, dan sampel dimasukkan ke dalam kelas yang outputnya memiliki perbedaan terkecil. Dalam model yang diusulkan untuk masalah klasifikasi dua kelas, nilai-nilai {1, +1} dan {0, +1} masing-masing dianggap sebagai nilai kelas, ketika fungsi hiperbolik dan logistik digunakan sebagai fungsi transfer output dari model yang diusulkan. Namun, dalam kasus menggunakan fungsi transfer linear untuk lapisan output dari model yang diusulkan mungkin lebih baik untuk menerapkan nilai {10, +10} atau {100, +100} sebagai nilai kelas. Nilai kelas yang lebih besar memperluas perbedaan kecil dalam output, membantu model menjadi lebih sensitif terhadap variasi input.

Perencanaan model hirarki untuk klasifikasi kelas ganda

Alasan untuk menggunakan pengklasifikasi hirarki fokus pada pengurangan kompleksitas. Dan mendeskripsikan pengklasifikasi hirarki sebagai bagian dari penggolong modular. Mereka menyarankan bahwa penggolong modular sering muncul ketika kombinasi faktor termasuk sejumlah besar kelas, kelas memiliki bentuk yang sulit (tidak kompak, cembung, atau terhubung), kelas tidak memiliki batas yang jelas, batas-batas sangat nonlinear, dan kesalahan klasifikasi beberapa poin membawa hukuman yang tinggi. [5] mendeskripsikan klasifikasi hirarki sebagai cara untuk mendeteksi data yang lebih sulit diklasifikasi untuk mengklasifikasikan data ini secara berbeda.

Dalam pendekatan ‘‘ satu lawan sisanya ’, untuk kasus kelas k, kelas dari kelas-kelas k ini pertama-tama dianggap sebagai kategori, dan kelas-kelas k 1 sisanya sebagai kategori lain, dan pengklasifikasi dua kelas dikonstruksi. Selanjutnya, kelas ini dikecualikan, dan kemudian proses yang dijelaskan diulang untuk kasus kelas k 1. Di sisi lain, kelas dari kelas k 1 yang tersisa dianggap sebagai kategori, dan sisanya k 2 = (k 1) 1 kelas sebagai kategori lain, dan classifier dua kelas kedua dibangun, dan seterusnya dan seterusnya sampai classifier dua kelas terakhir dibangun. Dengan cara ini, pengklasifikasi kelas dua k1 harus dibangun seluruhnya untuk kasus kelas k. Pendekatan ‘‘ one versus all ’’ mirip dengan pendekatan ‘‘ one versus rest ’dengan perbedaan sedikit. Dalam pendekatan ‘‘ satu lawan semua ’, untuk kasus kelas k, kelas dari kelas k ini juga dianggap sebagai kategori, dan kelas k 1 sisanya sebagai kategori lain, dan classifier twoclass dibangun; Namun, kelas ini tidak dikecualikan. Dengan cara ini, k classifier kelas dua harus dibangun seluruhnya untuk kasus kelas k.

Hasil Pembahasan

Perbandingan dengan model klasifikasi lainnya untuk benchmark dua kelas set data

Sebuah arsitektur yang terdiri dari dua input, tiga neuron tersembunyi dan satu output (N (2-3-1)) yang digunakan oleh Ripley (1994) telah ditemukan menjadi yang paling akurat di antara semua arsitektur JST lainnya, dengan tingkat kesalahan 9.4% . Namun, model yang kami usulkan mengungguli model ini pada bagian uji dari kumpulan data, dengan tingkat kesalahan sebesar 8,9%, peningkatan 5,32% dibandingkan dengan hasil jaringan syaraf tradisional terbaik sebesar 9,4%. Selain itu, berdasarkan sifat dari kumpulan data, diharapkan bahwa analisis diskriminan linear dan kuadratik tidak akan menjadi pengklasifikasi yang optimal seperti kelas-kelas ini. Setiap set data dibagi menjadi satu set pelatihan dan satu set tes, dan masing-masing model diterapkan dengan tepat. Tingkat kesalahan klasifikasi untuk setiap kelas dihitung dan disajikan, serta peningkatan persen dalam tingkat kesalahan untuk model yang diusulkan.

Multiple class data sets

Klasifikasi ke dalam beberapa kelas jauh lebih kompleks daripada klasifikasi dua kelas set data. Banyak metode klasifikasi menentukan partisi antara dua kelas set data. Kelas tambahan memerlukan penyesuaian pada metode klasifikasi yang sering menghasilkan tingkat kesalahan yang lebih tinggi. Untuk analisis diskriminan linear dan kuadratik, model n1 one-versus-all (di mana n adalah jumlah kelas) dan mengklasifikasikan setiap sampel seperti yang dijelaskan di atas.

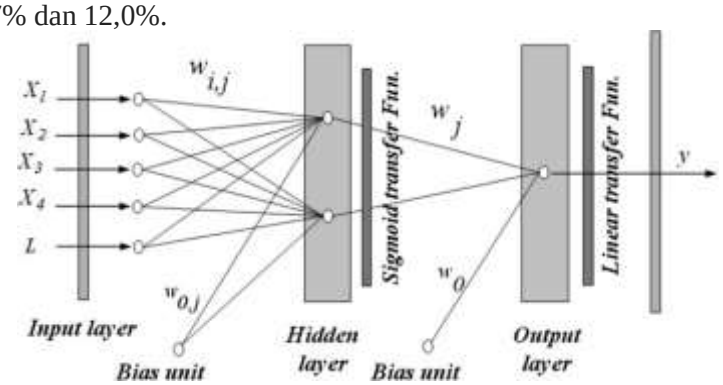
Fisher iris data set

Kumpulan data Fisher iris mungkin merupakan kumpulan data klasifikasi tertua dan paling banyak digunakan. Kumpulan data ini dinamai Fisher, yang menggunakannya dalam makalahnya yang semusim tahun 1936 tentang analisis diskriminan linear (Fisher, 1936). Set data terdiri dari 150 sampel, dibagi secara merata di antara tiga kelas. Kelas-kelas mewakili tiga spesies iris: Iris Setosa, Iris Versicolour, dan Iris Virginica. Setiap iris ditandai oleh empat atribut, (1) panjang sepal, (2) lebar sepal, (3) panjang kelopak, dan (4) lebar kelopak. Set data secara acak dibagi menjadi 75 pelatihan dan 75 sampel uji.

Gambar. 4. Struktur jaringan yang paling pas (set data Fisher iris), N (5-2-1).

Perbandingan dengan model-model klasifikasi lainnya untuk benchmark kelas set data ganda.

Dalam kasus klasifikasi data Fisher iris, mirip dengan bagian sebelumnya, beberapa arsitektur yang berbeda dari jaringan saraf tiruan telah dirancang dan diperiksa. Sebuah arsitektur yang terdiri dari empat input, dua neuron tersembunyi dan satu output (N (4-2-1)) sebagaimana diperiksa oleh Curram, Mingers, dan jaringan [16] telah ditemukan menjadi yang paling akurat di antara semua arsitektur jaringan saraf lainnya, dengan tingkat kesalahan 4,6% pada bagian uji dari kumpulan data. Namun, model yang diusulkan hirarki mengungguli model ini pada bagian uji dari kumpulan data, dengan tingkat kesalahan 1,3%, peningkatan 71,74%. Model yang diusulkan hirarki dan mesin vektor pendukung melakukan yang terbaik pada kedua pelatihan dan sampel uji dengan tingkat kesalahan 0,0% dan 1,3% masing-masing dalam pelatihan dan sampel uji, masing-masing. Analisis diskriminan kuadrat juga berkinerja terbaik dengan tingkat kesalahan 1,3% pada bagian uji dari kumpulan data (misclassifying hanya satu sampel); Namun, kinerjanya pada bagian pelatihan dari kumpulan data lebih buruk daripada model yang diusulkan hirarki dan mendukung model mesin vektor. Model yang diusulkan non-hirarki dan KNN keduanya memiliki tingkat kesalahan 2,7% pada bagian uji dari kumpulan data; Namun, tingkat kesalahan dari model yang diusulkan non-hirarki pada bagian pelatihan dari kumpulan data lebih baik daripada model KNN. Dengan cara yang sama, untuk KNN, semua atribut diskalakan oleh skor-z mereka sebelum menggunakan model KNN. Analisis diskriminan linear melakukan yang terburuk dalam pelatihan dan sampel uji dengan tingkat kesalahan masing-masing 10,7% dan 12,0%.



Gambar 1. Multi Layer NN

Dalam kasus klasifikasi data kaca Forensik, model usulan hirarki kami juga memiliki kesalahan terendah pada bagian pengujian dari kumpulan data dibandingkan dengan model-model lain yang digunakan untuk set data kaca Forensik, dengan tingkat kesalahan klasifikasi sebesar 26,8%. Seperti kasus sebelumnya, beberapa arsitektur jaringan syaraf tiruan yang berbeda telah dirancang dan diperiksa. Arsitektur berkinerja terbaik terdiri dari sembilan input, enam neuron tersembunyi dan satu output (N (9-6-1)) yang dirancang oleh Ripley (1994) untuk jaringan syaraf tiruan tradisional,

menghasilkan tingkat kesalahan 33,0%. Namun, kinerja ini adalah 18,79% lebih rendah dari model yang diusulkan hirarki. Selain itu, tingkat kesalahan klasifikasi untuk analisis diskriminan linear adalah 32,3%. Model yang diusulkan hirarki meningkatkan ini dengan 17,03%.
Set data kaca forensik menyajikan tantangan tambahan untuk analisis diskriminan kuadrat karena atribut komposisi barium dan besi hanya memiliki nilai nol untuk kelas-kelas tertentu, menghasilkan sarana dan varians nol. Kebalikan dari varians ini karenanya tidak terdefinisi. Karena fungsi diskriminan kuadrat membutuhkan matriks kovarian terbalik untuk setiap kelas, atribut ini harus dikecualikan untuk analisis diskriminan kuadrat. Analisis diskriminan kuadratik, dihambat oleh penghilangan dua atribut, misclassified 72,9% dari sampel, melakukan 63,24% lebih buruk daripada model yang diusulkan hirarki. KNN memiliki tingkat kesalahan terendah berikutnya dengan 29,2% salah diklasifikasi, 8,22% lebih tinggi dibandingkan dengan model yang diusulkan hirarki. Mesin vektor pendukung tingkat kesalahan 30,2%, 11,26% lebih tinggi dari model yang diusulkan hirarki. Untuk kedua klasifikasi klasifikasi data kelas ganda, model yang diusulkan hirarki bekerja lebih baik daripada jaringan saraf tiruan tradisional. Peningkatan bervariasi dari 71,74% menjadi 18,79% dibandingkan dengan jaringan saraf untuk data iris Fisher yang ditetapkan ke set data kaca Forensik. Selain itu, model yang diusulkan secara hirarki bekerja sebaik atau lebih baik daripada mendukung mesin vektor dan juga model klasifikasi tradisional lainnya seperti analisis diskriminan linier, analisis diskriminan kuadratik, dan KNN untuk kedua kumpulan data yang diperiksa. Hasil ini lagi menunjukkan bahwa model yang diusulkan menghasilkan hasil yang baik secara konsisten dalam berbagai kasus.

Kesimpulan

Klasifikasi memainkan peran penting dalam banyak aplikasi yang berkaitan dengan kecerdasan buatan dalam arti keputusan prediktif dalam pemrosesan informasi. Aplikasi ini mencakup berbagai bidang penelitian termasuk bisnis, kedokteran, biologi, pengenalan citra, penambangan data, dll. Banyak penelitian dalam klasifikasi telah menyatakan bahwa kinerja meningkat dalam model gabungan. Dalam model hibrida, tujuannya adalah untuk mengurangi risiko menggunakan model yang tidak pantas dengan menggabungkan beberapa model untuk mengurangi risiko kegagalan dan mendapatkan hasil yang lebih akurat. Biasanya, ini dilakukan karena proses yang mendasarinya tidak dapat ditentukan dengan mudah. Motivasi untuk menggabungkan model berasal dari asumsi bahwa salah satu tidak dapat mengidentifikasi proses pembuatan data yang benar atau bahwa model tunggal mungkin tidak cukup untuk mengidentifikasi semua karakteristik dari rangkaian waktu.

Dalam makalah ini, model hibrida baru dari jaringan saraf tiruan diusulkan sebagai model alternatif untuk masalah klasifikasi menggunakan model regresi linier berganda. Tujuan utama dari model yang diusulkan adalah menggunakan keuntungan unik dari model regresi linier berganda dalam pemodelan linier untuk mengatasi kekurangan pemodelan linear dari jaringan saraf tiruan tradisional. Model yang diusulkan terdiri dari dua fase, (i) meringkas komponen linier dalam atribut dalam atribut baru untuk pemodelan yang lebih baik dengan jaringan saraf, dan (ii) mengklasifikasi data dengan jaringan saraf menggunakan atribut asli dan atribut linier yang dihasilkan oleh beberapa linier regresi. Enam tolok ukur terkenal (sintetis dan kehidupan nyata) dan kumpulan data dunia nyata — kumpulan data sintetis Ripley, set data Pima Indian Diabetes, kumpulan data Fisher iris, set data kaca Forensik, kumpulan data kredit Jepang, dan set data ekspresi gen - digunakan dalam makalah ini untuk menunjukkan kelayakan dan keefektifan model yang diusulkan untuk tugas klasifikasi dua kelas dan beberapa kelas. Hasil yang diperoleh dari masalah dua kelas menunjukkan bahwa model yang diusulkan menjadi lebih unggul untuk semua model alternatif untuk kedua set data benchmark sintetis dan kehidupan nyata.

Untuk menyelesaikan masalah multi-kelas, dalam makalah ini versi hirarki dari model yang diusulkan dikembangkan dengan memeriksa tiga pendekatan yang berbeda termasuk " satu lawan satu ", " satu lawan istirahat ", dan " satu lawan semua ' '. Di antara pendekatan ini, pendekatan ‘‘ satu banding semua ’menghasilkan hasil yang lebih akurat dan mengajukan permohonan untuk membangun versi hirarki dari model yang diusulkan. Hasil empiris untuk kelompok masalah ini

menunjukkan bahwa model hirarki yang diusulkan secara konsisten mengungguli perceptrons multilayer tradisional dan model lain yang digunakan dalam makalah ini seperti analisis diskriminan linear, analisis diskriminan kuadratik, tetangga K-terdekat, dan mesin pendukung vektor.

DAFTAR PUSTAKA

[1] Chakraborty, S. (2009). Simultaneous cancer classification and gene selection with Bayesian nearest neighbor method: An integrated approach. Computational Statistics and Data Analysis, 53, 1462–1474.
[2] Amasyali, M., & Ersoy, O. (2008). Cline: A new decision-tree family. IEEE Transactions on Neural Networks, 19(2), 356–363.
[3] Banerjee, A., Kiran, K., Murty, U., & Venkateswarlu, Ch. (2008). Classification and identification of mosquito species using artificial neural networks. Computational Biology and Chemistry, 32, 442–447.
[4] Acharya, U., Bhat, P., Iyengar, S. S., Rao, A., & Dua, S. (2003). Classification of heart rate data using artificial neural network and fuzzy equivalence relation. Pattern Recognition, 36, 61–68.
[5] Aci, M., Inan, C., & Avci, M. (2010). A hybrid classification method of k nearest neighbor Bayesian methods and genetic algorithm. Expert Systems with Applications, 37, 5061–5067.
[6] Amanda, J. C. (1999). Combining artificial neural nets: Ensemble and modular multinet systems. London: Springer.
[7] Asuncion, A., & Newman, D. (2007). UCI machine learning repository. Irivine, CA:
[8] University of California, School of Information and Computer Science, 2007.
[9] D. C. Sulaiman and T. M. S. Mulyana, "Web-Based Writing Learning Application of Basic Hanacaraka Using Convolutional Neural Network Method," *Ultimatics : Jurnal Teknik Informatika*, pp. 28–34, Jun. 2023, doi: 10.31937/ti.v15i1.2993.
[10] M. Freddy and T. M. S. Mulyana, "Determining Computer Opponent's Actions in Strategy Game Using K-Nearest Neighbour Algorithm," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 3, Dec. 2022, doi: 10.28932/jutisi.v8i3.5137.
[11] T. Matius and S. Mulyana, "SEGMENTASI CITRA MENGGUNAKAN HEBB-RULE DENGAN INPUT VARIASI RGB," *Jurnal Teknologi Informasi*, vol. 11, no. 1, Juni 205, pp. 34–443, 2015.
[12] Billings, S., & Lee, K. (2002). Nonlinear Fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm. Neural Networks, 15(2), 262–270.
[13] Breiman, L. (1999). Prediction games and arcing algorithm. Neural Computation, 11, 1493–1517.
[14] Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., et al. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences of the United States of America 97(1), 262–267.
[15] Castellani, M., & Rowlands, H. (2009). Evolutionary artificial neural network design and training for wood veneer classification. Engineering Applications of Artificial Intelligence, 22, 732–741.
[16] Chaovalitwongse, W. (2007). On the time series k-nearest neighbor classification of abnormal brain activity. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 37(6).
[17] Chen, S., Lin, S., & Chou, S. (2010). Enhancing the classification accuracy by scattersearch-based ensemble approach. Applied Soft Computing xxx, xxx–xxx, 2010.
[18] Christianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines. Cambridge University Press.