MENGUJI FUNGSI K-NN DAN FUNGSI NEURAL-NET PUSTAKA BAHASA R UNTUK KLASSIFIKASI DATA

Suarga
Program Studi Teknik Informatika,
STMIK Dipanegara, Makassar
suarga@dipanegara.ac.id

Abstrak

Klassifikasi adalah metoda pembelajaran mesin terbimbing dimana data digunakan sebagai "data latih" untuk membangun model klassifikasi, kemudian model diterapkan pada "data uji" untuk verifikasi ketepatan prediksi-nya. Ada dua metoda yang diperiksa pada paper ini, k-Nearest Neighbor (k-NN) dan Artificial Neural Network (ANN), untuk melakukan klassifikasi dari tiga kumpulan data. Hasil yang diperoleh akan dibandingkan dan ketepatan prediksi akan dinyatakan.

Kata kunci : Data Mining, Data Science, Supervised Machine Learning, Klassifikasi

Abstract

Classification is a supervised learning method where a classified data set is used as a" training data" to build the classification model, and then the model is applied to a " test data set" to verify prediction accuracy. Two methods was explored in this paper, k-Nearest Neighbor (k-NN) and Artificial Neural Network (ANN) to classify three different data sets. The results was compared and the accuracy of both methods was declared.

Keywords: Data Mining, Data Science, Supervised Machine Learning, Classification

1. PENDAHULUAN

Klassifikasi merupakan salah satu metoda penentuan kelas dari suatu objek data ke dalam kelas tertentu yang tersedia. Ada dua langkah utama dalam proses klassifikasi yaitu: (1) membangun model sebagai prototipe dari sejumlah kelas data dan (2) menggunakan model untuk menentukan / memprediksi kelas dari data lain yang sejenis. Klassifikasi merupakan salah satu topik dalam Data Mining atau Data Science.

Teknik klassifikasi telah digunakan pada berbagai aplikasi, seperti klassifikasi hewan berdasarkan sejumlah atribut yang dimilikinya, klassifikasi penyakit kanker kulit (melanomia), klassifikasi dari sel darah merah apakah normal atau mengandung impuritan, dan sebagainya.

Beberapa teknik klassifikasi telah ditemukan oleh para ahli, seperti k-Nearest Neighbor (k-NN), Artificial Neural Network (ANN, Neural net), Support Vector Machine (SVM), Naïve Bayes, dan Decision Tree [Chiu,2015] [Ledolter,2013] [Lantz,2015] [Tomey, 2014] [Mayor, 2015], namun hanya dua teknik klassifikasi yang diselidiki pada paper ini yaitu k-NN dan Neural-net, dan kedua-nya tersedia sebagai pustaka fungsi dalam bahasa-R.

Proses pengujian pada metoda "Supervised Learning" [Tomey, 2014] [Usuelli,2014] dilakukan dengan membagi dua setiap jenis data-set menjadi *data latih (training data)* dan *data uji (testing data)*. Data latih pada percobaan ini dipilih secara acak sekitar 90% populasi data, dan sisanya digunakan sebagai data uji. Ada tiga jenis data set yang digunakan, yaitu *data bunga Iris*,

data Forensic Glass, dan *German Credit Data*. Ketiga data ini juga telah digunakan oleh beberapa peneliti sebelumnya [Ledolter,2013] [Lantz, 2015] [Daroczi, 2015] [Mayor, 2015].

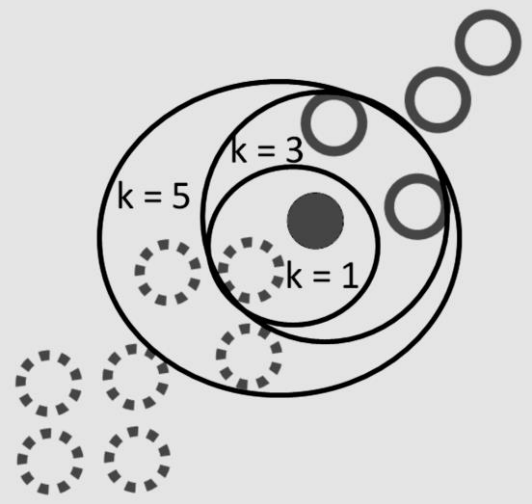
Hasil eksplorasi disajikan dalam bentuk tabel yang menyatakan teknik klassifikasi, data set, serta prosentase data yang diprediksi benar.

2. METODA PENELITIAN

Ada dua metoda klassifikasi yang akan diuji dalam ekplorasi ini yaitu k-Nearest Neighbor (k-NN) dan Artificial Neural Network (Neural-net), keduanya tersedia sebagai pustaka bahasa R. Bahasa R adalah bahasa untuk analisa-data utamanya analisa statistik dan data mining. Bahasa R diciptakan oleh dua orang professor Statistik, George Ross Ihaka dari Universitas Auckland di New Zealand, dan Robert Clifford Gentlemen dari Canada, kini ditangani oleh R-Development Core Team dari proyek GNU. Perangkat lunak R gratis dan diambil dari situs http://www.r-project.org/.

2.1. k-Nearest Neighbor (k-NN)

Teknik k-NN, adalah salah satu teknik klassifikasi data yang memanfaatkan jarak data ke k-buah data tetangga terdekat (K-Nearest Neighbor). Illustrasi k-NN disajikan pada Gambar-1 berikut ini.



Gambar 1 : Tetangga terdekat dengan k=1, 3, dan 5 [Mayor, 2015]

Formula jarak yang digunakan pada umumnya adalah “Euclidean Distance” yang merupakan akar dari jumlah kuadrat selisih setiap atribut data.

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (1)$$

Data yang diprediksi merupakan anggota kelas dari k-data tetangga dengan jarak terdekat darinya. Klassifikasi k-NN dalam R dapat diperoleh dari pustaka sebagai berikut:

```
a. library(class)
   model <- knn(datatrain, datatest, class, k=n)
```

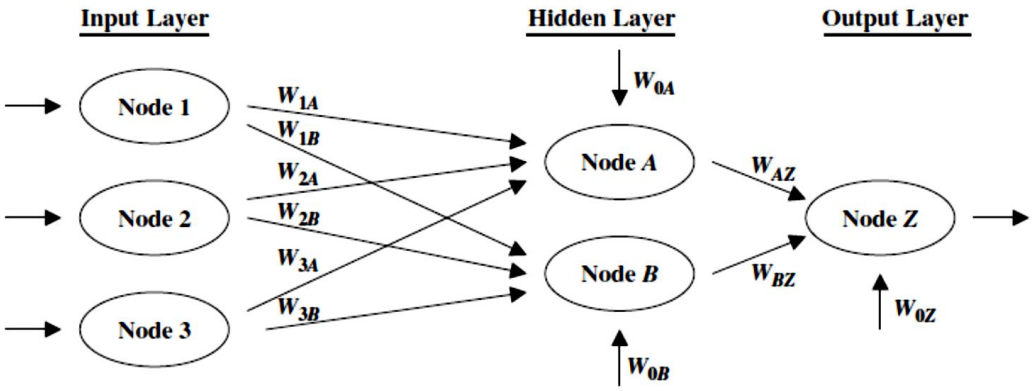
model merupakan hasil prediksi dari klassifikasi datatest

```
b. library(caret)
   model <- train(datatrain, class, method='knn')
   hasil <- predict(model, newdata=datatest) #fungsi prediksi
```

2.2. Artificial Neural Network (ANN , Neural-net)

Neural-net merupakan suatu konsep rekayasa pengetahuan dalam bidang kecerdesan buatan (Artificial Intelligence, AI) yang dirancang dengan mengadopsi sistem saraf manusia. Bagian terkecil dari saraf otak manusia disebut neuron, dan terdapat sekitar 10 milyar neuron didalam otak manusia.

Sebuah neuron terdiri atas badan sel (soma), sejumlah serat (dendrit) yang menyalurkan informasi ke soma, dan sebuah serat tunggal (axon) yang menyalurkan informasi keluaran dari neuron ke sel berikutnya. Setiap neuron terhubung ke neuron lainnya melalui dendtrit dan axon sehingga neuron-neuron dapat bekerja secara paralel berkecepatan tinggi.



Gambar 2 : Model-ANN (Neural-net)

Bahasa R menyediakan fungsi untuk pelatihan memakai Neural-net yaitu :

```
a. library(nnet)
   model < nnet(formula,data=train, size=hidden, maxit=n)
b. library(neuralnet)
   model <- neuralnet(formula, data=train, hidden=n, stepmax=m)
c. hasil <- predict(model, newdata=test)
```

2.3. Data Set

Data set yang dipersiapkan untuk menguji kedua teknik klassifikasi tersebut diatas ada tiga macam yaitu, data bunga Iris, data Forensic Glass, dan German Credit Data. Ketiga data set ini merupakan data yang telah memiliki klassifikasi pada setiap baris data-nya, sehingga cocok untuk dijadikan data latih dan data uji dalam supervised-learning.

2.3.1. Data Bunga Iris

Data Iris merupakan data klassifikasi dari tiga jenis bunga Iris. Data ini dapat diperoleh dari UCI Machine Learning Repository (https://archive.ics.uci.edu). Data Iris terdiri atas 150 observasi masing-masing dengan 4 atribut/fitur (Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) dan 1 klassifikasi (Species). Ada 50 data jenis Iris-Setosa, 50 data Iris-Versicolor, dan 50 data Iris-Virginica. Pada percobaan ini 130 data bunga Iris yang dipilih secara acak akan digunakan sebagai data latih, dan sisanya 20 data sebagai data uji.

2.3.2. Data Forensic Glass

Data Forensic Glass merupakan klassifikasi dari beberapa jenis kaca/gelas. Data ini tersedia dalam paket R melalui library(MASS), dan diberi nama "*fgl*".

Terdapat 214 observasi dalam data Forensic Glass, dengan 9 atribut/fitur dan 1 kelas (type). Fitur merupakan kandungan beberapa mineral dalam gelas, seperti Na, Mg, Al, dan sebagainya. Ada 6 macam klassifikasi gelas, yaitu WinF (70 data), WinNF(76 data), Veh(17 data), Con(13 data), Tabl (9data), Head (29 data). Dalam percobaan ini hanya 7 fitur yang digunakan yaitu: RI, Na, Mg, Al, Si, K, dan Ca. Fitur ke-8 yaitu Ba, dan fitur ke-9 yaitu Fe, tidak disertakan dalam analisa karena nilai-nya sebagian besar = 0.0.

2.3.3. German Credit Data

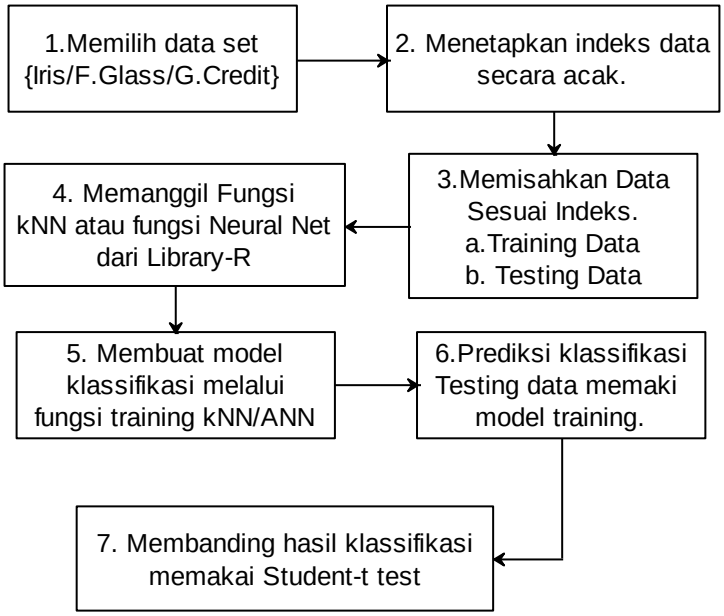
Data set German Credit bersumber dari UCI (University of California at Irwin) Machine Learning Repository (https://archive.ics.uci.edu). Data ini terdiri atas 1000 record , dan 21 kolom, namun tidak semua kolom dapat dijadikan fitur data klassifikasi.

Pada percobaan ini untuk keperluan klassifikasi jumlah kolom direduksi menjadi 5 saja : yaitu dipilih variabel/kolom: "Default","Duration","Amount","Installment","Age". Kolom 1 (Default) di-gunakan sebagai kelas-nya, dimana Default=0 menyatakan gagal memperoleh kredit dan Default=1 berhasil memperoleh kredit, kolom lainnya sebagai fitur dari klassifikasi.

2.4. Algoritma Pengujian

Pengujian metoda klassifikasi k-NN dan Neural-net dilakukan mengikuti langkah berikut ini:

1. Siapkan data, (Iris / Forensic Glass / German Credit)
2. Tetapkan indeks secara acak dari data untuk digunakan sebagai indeks training data
3. Training_data = data[indeks] # sesuai indeks acak, testing_data = data[-indeks] # sisa-nya
4. Panggil library fungsi klassifikasi (knn atau neuralnet)
5. Lakukan training memakai fungsi klassifikasi
6. Lakukan prediksi klassifikasi dari testing_data
7. Hitung akurasi hasil prediksi dengan membandingkan data asli, lalu menguji perbedaan hasil kNN dan ANN melalui uji student-t.



Gambar 3 : Blok Struktur Metoda Pengujian

Tiga buah program R disusun berdasarkan algoritma diatas, masing-masing adalah program untuk data Iris, program untuk data Forensic Glass, dan program untuk German Credit data. Pada setiap program dilakukan dua kali klassifikasi, masing-masing memakai metoda k-NN dan metoda Neural-net.

3. HASIL DAN PEMBAHASAN

Klassifikasi memakai teknik k-NN dan Neural-net telah diterapkan terhadap tiga macam data melalui program bahasa R sesuai dengan algoritma pengujian. Selanjutnya akurasi yang diperoleh pada setiap data-set dan setiap metoda klassifikasi dibandingkan.

Berikut ini disajikan tabel hasil pengujian kedua metoda klassifikasi terhadap tiga kumpulan data tersebut.

Tabel 1 : Tabel perbandingan hasil percobaan

Klassifikasi	Data Iris	Data Forensic Glass	German Credit
#data latih	130	189	900
#data uji	20	25	100
Rasio data uji	15.38%	13.22%	11.11%
#Fitur	4	7	4
k-Nearest Neighbor	100% (k=3)	88% (k=3)	100% (k=3)
Neural network	100% (hidden=5)	56% (hidden=7)	73% (hidden=5)
Kesimpulan	kNN = NNet	kNN > Nnet	kNN > NNet

Klassifikasi data Iris dengan 4 fitur berhasil diprediksi dengan sempurna, kedua metoda memberikan hasil klassifikasi benar 100%. Data Forensic Glass dengan 7 fitur tidak dapat di-prediksi hingga 100%, metoda k-NN memberikan akurasi 88% , metoda neural-net hanya 56%. Data German Credit dengan 4 fitur dapat diprediksi 100% memakai k-NN tetapi hanya 73% memakai Neural-net.

Perbedaan antara hasil dari metoda k-NN dengan hasil metoda Neural-net telah diuji dengan test perbedaan student-t test dengan null-hypothesis (H0) bahwa keduanya sama. Hasilnya: t = 1.4654, p-value=0.2609, atau H0 harus ditolak pada alpha=0.95, dengan kata lain H1 diterima bahwa kedua metoda ini memberikan hasil yang berbeda.

Jumlah fitur yang digunakan dalam training mempengaruhi hasil klassifikasi terutama pada metoda Neural-net, karena semakin banyak fitur berarti jumlah neuron input pada model akan semakin banyak, demikian pula hidden neuron-nya. Tidak ada petunjuk umum tentang jumlah neuron pada hidden-layer, maka pada percobaan ini dipilih jumlah hidden neuron mendekati jumlah fitur.

Diduga bahwa jumlah fitur pada data menyebabkan data Forensic Glass dengan 7 fitur hasil prediksinya ternyata tidak dapat mencapai 100%.

Hal lain yang terlihat dari eksperimen ini adalah metoda k-NN terlihat lebih unggul dalam prediksi klassifikasi dibanding metoda Neural-net.

4. KESIMPULAN DAN SARAN

Metoda k-Nearest Neighbor (k-NN) dan Artificial Neural Network (Neural net) merupakan dua metoda dari beberapa metoda yang ada untuk klassifikasi. Kedua metoda tersebut telah diuji coba terhadap tiga jenis data, data bunga Iris, data Forensic Glass, dan data German Credit. Kesimpulan sementara yang dapat diambil adalah sebagai berikut:

1. Kedua metoda berhasil melakukan klassifikasi walaupun tidak semua akurat 100%
2. Jumlah fitur pada data mempengaruhi prediksi, terlihat data dengan fitur lebih banyak (Forensic Glass) diprediksi lebih rendah dari data yang lain.
3. Pada percobaan ini metoda k-NN terlihat lebih unggul dari metoda Neural-net.

Percobaan selanjutnya yang dapat dilakukan adalah dengan membandingkan kedua metoda ini dengan metoda klassifikasi yang lain, seperti Naïve-Bayes, dan Decision-tree. Serta menggunakan data-set yang lebih banyak fitur-nya. Selain itu variasi jumlah k pada k-NN dan variasi jumlah hidden-neuron pada Neural-net dapat dibandingkan.

DAFTAR PUSTAKA

Chiu, David., 2015, "*Machine Learning with R Cookbook",* Packt Publishing, Open Source, Birmingham.

Daroczi, Gergely., 2015, "*Mastering Data Analysis with R*", Packt Publishing, Open Source, Birmingham.

Grolemund, G.,2014, "*Hands-On Programming with R*", O'Reilly Media.Inc., Sebastopol CA.

Lantz, Brett., 2015, "*Machine Learning with R",* 2nd Ed, Packt Publishing, Open Source, Birmingham.

Ledolter, Johannes., 2013, "*Data Mining and Business Analytics with R*", John Wiley & Sons, Inc., New Jersey.

Mayor,Eric., 2015, "*Learning Predictive Analytics with R",* Packt Publishing, Open Source, Birmingham.

Tomey, Dan., 2014, "*R For Data Science*", Packt Publishing, Open Source, Birmingham.

Usuelli, Michele., 2014, "*R Machine Learning Essentials",* Packt Publishing, Open Source, Birmingham.

Zumel, N., and Mount,J., 2014, "*Practical Data Science with R*", Manning Publications Co. Shelter Island, NY.