NASKAH PUBLIKASI

ANALISIS PERBANDINGAN PADA METODE PENGHITUNGAN JARAK ANTAR DATA PADA ALGORITMA K-NN DAN LVQ UNTUK KLASIFIKASI DATA

Program Studi Informatika

Disusun oleh:

Hari Nugraha

5160411129

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI DAN ELEKTRO UNIVERSITAS TEKNOLOGI YOGYAKARTA

2020

NASKAH PUBLIKASI

ANALISIS PERBANDINGAN PADA METODE PENGHITUNGAN JARAK ANTAR DATA PADA ALGORITMA K-NN DAN LVQ UNTUK KLASIFIKASI DATA

Disusun oleh

Hari Nugraha

5160411129

Pembimbing

Muhammad Fachrie, S.T., M.Cs.

Tanggal : 25-02-2020.

ANALISIS PERBANDINGAN PADA METODE PENGHITUNGAN JARAK ANTAR DATA PADA ALGORITMA K-NN DAN LVQ UNTUK KLASIFIKASI DATA

Hari Nugraha
Program Studi Teknik Informatika, Fakultas Teknologi Informasi dan Elektro, Universitas Teknologi Yogyakarta Jl. Ringroad Utara Jombor Sleman Yogyakarta E-mail : hari.nugraha@student.uty.ac.id

ABSTRAK

Klasifikasi adalah suatu teknik yang digunakan untuk membangun model klasifikasi dari sampel data pelatihan. k-*NN* dan LVQ merupakan algoritma klasifikasi yang memiliki parameter penting yang mempengaruhi kinerjanya. Parameter tersebut adalah nilai metode perhitungan jarak. Jarak antara dua titik data ditentukan oleh perhitungan matriks jarak sebelum dilakukan proses klasifikasi oleh kedua algoritma tersebut. Tujuan penelitian ini adalah untuk menganalisis dan membandingkan akurasi kinerja k-*NN* dan LVQ menggunakan fungsi jarak *Euclidean Distance, Manhatan Distance, Minkowski Distance, Canberra Distance, Cosine Distance,* dan *Chebishev Distance* berdasarkan sudut pandang akurasi. Adapun data yang digunakan adalah dataset yang bersumber dari *Kaggle.com*. Metode evaluasi yang digunakan adalah *k-Fold Cross Validation.* Hasil analisis menunjukan bahwa *Canberra Distance* memiliki performa lebih baik pada kedua algoritma dengan rata-rata akurasi sebesar 72.53%. *Manhattan Distance* memiliki performa lebih baik pada algoritma k-*NN* dengan rata-rata akurasi sebesar 76.76%. Sedangkan *Cosine Distance* memiliki performa lebih baik pada algoritma LVQ dengan rata-rata akurasi 69.84%.

Kata kunci: Klasifikasi, Algoritma, Perhitungan Jarak

1. PENDAHULUAN

Perkembangan kecerdasan buatan dalam teknologi khususnya *machine learning* dan *data mining* sangatlah cepat. *Machine learning* atau pembelajaran mesin adalah salah satu cabang dari kecerdasan buatan yang berkembang pada sistem dan dapat melakukan pembembelajaran sendiri. Hingga saat ini perkembangan *machine learning* banyak digunakan untuk membantu pekerjaan manusia. Salah satu Teknik dalam *machine learning* adalah dengan mempelajari kedekatan jarak antar data. Sekurangnya ada dua algoritma *machine learning* yang mengunakan pendekatan jarak antar data yaitu *K-Nearest Neighboar* dan *Learnng Vector Quantization.*

Untuk mendapatkan hasil klasifikasi yang optimal maka kedua algoritma di atas baik k-NN ataupun LVQ cukup bergantung terhadap metode perhitungan jarak yang digunakan. Banyaknya metode perhitungan jarak membuat peneliti memiliki banyak alternatif. Di sisi lain peneliti harus mencoba metode satu persatu untuk mendapatkan yang terbaik yang tentunya akan memakan waktu lama.

Berdasarkan permasalahan tersebut penulis mengusulkan untuk menganalisis Perbandingan pada Metode Penghitungan Jarak Antar Data pada Algoritma k-*NN* dan LVQ untuk Klasifikasi Data.

Adapun batasan masalah dari penelitian ini adalah :

a. Analisis hanya untuk mengetahui metode perhitungan jarak yang terbaik.
b. Metode perhitungan jarak pada k-*NN* dan LVQ yang dibandingkan hanya enam, yaitu *Euclidean Distance, Manhattan Distance, Minkowsky Distance, Canberra Distance, Cosine Distance,* dan *Chebyshev Distance*.
c. Aplikasi berjalan pada *platform console*.
d. *Dataset* yang digunakan adalah *Iris*

Flower, Email Spam, Glass, Heart Desise, dan *Titanic.*

Algoritma k-*NN* dan LVQ memiliki parameter penting pada kinerjanya. Salah satu parameter tersebut adalah perhitungan jarak antar data. Perhitungan jarak antar data menjadi faktor penting yang bergantung pada kumpulan data yang dapat mempengaruhi kinerja algoritma. Oleh karena itu diperlukan pengukuran dan perbandingan kinerja metode perhitungan jarak pada proses klasifikasi dari sudut pandang akurasi untuk mendapatkan metode perhitungan jarak yang optimal untuk algoritma tertentu.

2. LANDASAN TEORI

2.1. *K-Nearest Neighbor* (k-*NN*)

K-Nearest Neighbor (k-*NN*) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut [9]. Sedangkan menurut Kusrini dkk, algoritma *k-Nearest Neighbor* adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama dengan berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada yang memiliki kesamaan (*similiarity*) [4]. Tujuan dari algoritma ini untuk mengklasifikasikan objek baru berdasarkan atribut dan *training sample*. *Classifier* tidak menggunakan model apapun untuk dicocokan dan hanya berdasarkan pada memori.

2.2. Learning Vector Quantization

LVQ adalah jaringan single layer yang terdiri dari dua lapisan yaitu lapisan input dan output [7]. Menurut Kusumadewi, Learning Vector Quantization (LVQ) adalah suatu metode untuk melakukanpembelajaran pada lapisan kompetitif yang terawasi [5]. Suatu lapisan kompetitif akan secara otomatis belajar untuk mengklasifikasikan vektor-vektor input . Kelas-kelas yang didapatkan sebagai hasil dari lapisan kompetitif ini hanya tergantung pada jarak antara vektor-vektor input . Jika 2 vektor input mendekati sama, maka lapisan kompetitif akan meletakkan kedua vektor input tersebut kedalam kelas yang sama.

2.3. Metode Perhitungan Jarak

Pada tahapan algoritma k-*NN* dan algoritma LVQ baik keduanya sama- sama melakukan analisis perhitungan jarak antar data, karena itu diperlukan metode perhitungan lebih dari satu agar dapat kita temukan perbandingan diantara beberapa metode yang diteliti. Berikut metode perhitungan jarak yang akan dianalisi dalam penelitian ini.

1. *Euclidean Distance*

Euclidean Distance merupakan runus perhitungan jarak antar data yang perhitungannya menggunakan konsep *Pythagoras* [6].

$$(,) = \sqrt{\sum(\quad -)^2} \qquad (2.1)$$

2. *Manhattan Distance*

Manhattan Distance merupakan runus perhitungan jarak antar data yang perhitungannya dengan cara menjumlahkan semua selisih dari jarak [6].

$$(,) = \sum| \quad - | \qquad (2.2)$$

3. *Minkowsky Distance*

Minkowsky Distance merupakan runus perhitungan jarak antar data yang perhitungannya menggunakan konsep aljabar dengan objek vektor berdimensi n dan r bukan 1 dan 2 [6].

$$(,) = (\sum| \quad - |) \qquad (2.3)$$

4. *Canberra Distance*

Matriks jarak *Canberra Distance* digunakan ketika diperlukan untuk mendapatkan jarak dari pasangan titik dimana data tersebut berupa data asli dan berada dalam ruang vektor [9].

| − |

= (Σ) (2.4)

5. Cosine Distance

Cosine Distance dapat diimplementasikan untuk menghitung nilai jarak antar data dan menjadi salah satu teknik untuk mengukur kemiripan teks yang popular [3].

Σ (2.5)

6. Chebishev Distance

Chebishev Distance dapat disebut juga jarak nilai maksimum. Metode ini mencari besarnya absolut dari perbedaan antara koordinat sepasang objek. Perhitungan jarak ini dapat digunakan untuk variabel ordinal dan kuantitatif [6].

(2.6)

2.4. Datasets

Dalam sebuah penelitian khususnya *machine learning* maka proses pembelajaran dapat dilakukan jika telah dimiliki data atau sering disebut sebagai *dataset.* Jika dimiliki 150 *instance* pada *dataset* maka harus dibagi menjadi dua. Data yang digunakan pada tahap *training* akan disebut dengan istilah *training set*. Sedangkan yang digunakan pada tahap pengujian disebut *test set* [2].

a. Data latih

Data latih adalah sebagian besar data yang diambil dari sebuah *dataset* untuk melakukan pelatihan oleh algoritma tertentu agar nantinya dapat menentukan kelas dari data yang akan diujikan.

b. Data uji

Data uji adalah sebagian kecil data yang diambil dari *dataset* untuk dilakukan pengujian oleh algoritma tertentu. Algoritma yang dimaksud tentunya membutuhkan *data latih* untuk menjadi landasan algoritma tersebut dalam mengklasifikasi. Dengan melakukan pengujian ini kita dapat mengetahui seberapa besar akurasi yang didapat oleh algoritma tertentu.

2.5. Confusion Matrix

Confusion Matrix adalah sebuah konsep dari teknik *machine learning*. *Confusion Matrix* memiliki informasi tentang data aktual dan hasil prediksi sebuah klasifikasi yang telah dilakukan oleh metode klasifikasi. *Confusion Matrix* mempunyai dua dimensi yaitu dimensi yang berisi data aktual suatu objek dan dimensi yang berisi hasil prediksi Teknik klasifikasi. Tabel 2.2 adalah contoh ilustrasi *confusion matrix* [8].

Tabel 2. 1 *Confusion Matrix*

		True values	
Prediction	*True*	*TP Correct result*	*FP Unexpected result*
	False	*FN Missing result*	*TN Correct absence of result*

Pada confusion matrix terdapat beberapa istilah yang digunakan pada kasus klasifikasi yaitu :

a. *True Positive* (TP) : data positif yang terdeteksi benar
b. *False Positive* (FP) adalah data negatif namun terdeteksi dengan benar
c. *False Negative* (FN) adalah data positif yang terdeteksi sebagai data negatif
d. *True Negative* (TN) adalah data negatif yang terdeteksi benar.

Beberapa perhitungan kinerja klasifikasi dapat dijelaskan dari confusion matrix. Berikut perhitungan kinerja klasifikasi tersebut adalah akurasi. akurasi adalah presentase dari jumlah total prediksi yang benar pada proses klasifikasi [1].

= + (2.7)

3. METODOLOGI PENELITIAN

3.1. Data Penelitian

Obyek penelitian yang berupa metode perhiitungan jarak pada kedua algoritma tersebut memerlukan dataset untuk mendapatkan hasil akurasi sebagai tolak ukur perbandingan. *Dataset* yang digunakan dalam penelitian ini menggunakan lima jenis data yang diambil dari *repositori Kaggle* (https://www.kaggle.com). Berikut lima jenis *dataset* yang digunakan dalam penelitian ini.

a. Dataset Iris Flower

Dataset ini memiliki 150 data yang nantinya akan dibagi menjadi dua yaitu 100 untuk data latih dan 50 untuk data uji. Selain itu dataset ini memiliki 4 *fiture* dan target klasifikasinya terbagi menjadi tiga kelas yaitu *Iris-Setosa, Iris-Versicolor,* dan *Iris-Virginica.*

b. *Dataset Glass*

Dataset ini memiliki 214 data dari *Glass* yang terbagi menjadi tujuh tipe. *Dataset* ini memiliki sembilan *fiture* yang nantinya tidak perlu di *filter* lagi.

c. Analisis pada Dataset Email Spam

Dataset ini memiliki 1000 data email yang terbagi menjadi email yang termasuk spam atau tidak. *Dataset* ini memiliki 57 *fiture* yang nantinya tidak perlu di *filter* lagi.

d. Dataset Heart Desiase

Dataset ini memiliki 303 data yang mengacu pada adanya penyakit jantung atau tidak. *Dataset* ini memiliki tiga belas *fiture* yang nantinya tidak perlu di *filter* lagi.

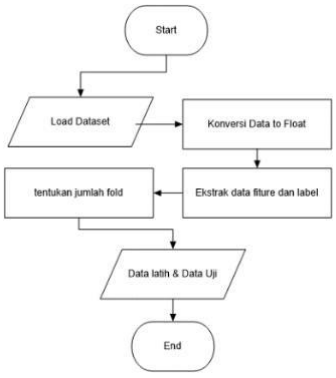
e. Dataset Dataset Titanic

Dataset ini memiliki 892 data kelas yaitu mengacu pada selamatnya penumpang atau tidak saat kecelakaan kapal, *Dataset* ini memiliki enam belas *fiture* yang nantinya tidak perlu di *filter* lagi.

3.2. Metode Penelitian a.

Alur Pemisahan Dataset

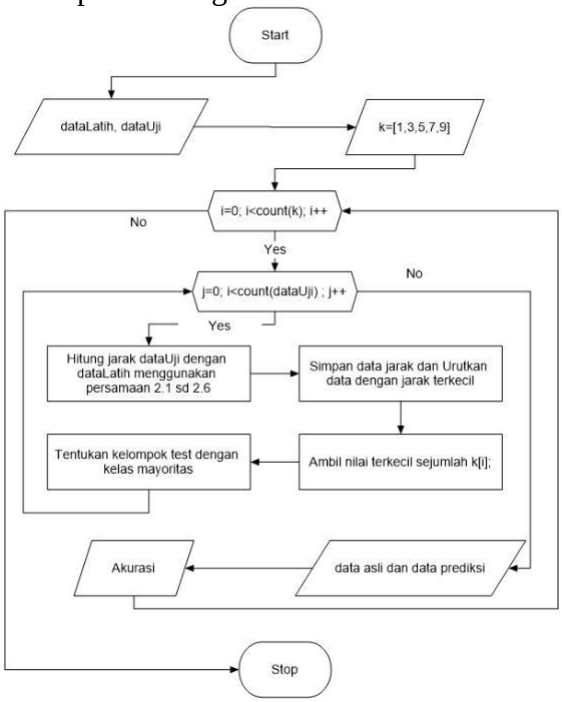
Dataset yang telah diambil dari repositori Kaggle perlu dilakukan pembagian sub dataset yaitu data latih dan data uji menggunakan metode k-*Fold Cross Validation*.



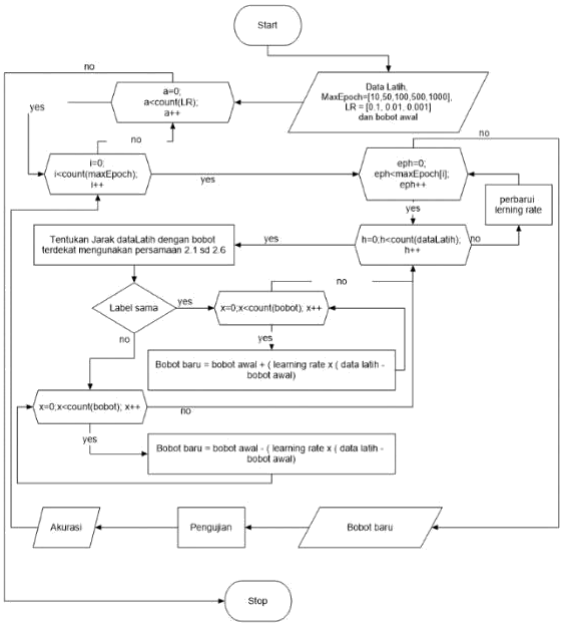
Gambar 1 Alur Pemisahan Data

b. Perancangan Sistem

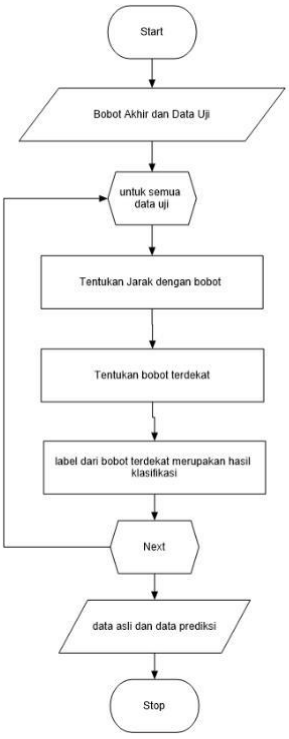
Perancangan sistem merupakan proses membuat gambaran sistem yang akan dibangun. Gambaran sistem ditampilkan dengan flowchart



Gambar 2 Flowchart algoritma k-NN



Gambar 3 Flowchart algoritma LVQ



Gambar 4 Flowchart sub pengujian LVQ

Gambar 5 Flowchart perhitungan akurasi

4. HASIL DAN PEMBAHASAN

4.1. Hasil

a. Perbandingan Secara Umum

Tabel 1 Perbandingan Secara Umum

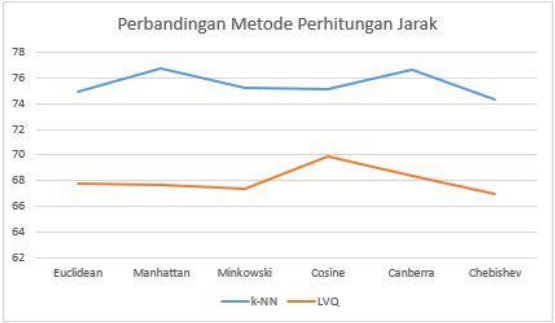
Mtd	Dataset										Avg
	Iris Flower		*Glass*		Email Spam		*Heart*		*Titanic*		
	kNN	LVQ	kNN	LVQ	kNN	LVQ	kNN	LVQ	kNN	LVQ	
Euc	97.01	91.00	62.27	35.89	74.26	69.00	61.92	63.31	78.99	79.61	71.33
Man	96.53	89.16	**63.86**	31.47	**77.96**	**72.01**	**66.47**	70.22	79.00	75.63	72.23
Min	**97.33**	91.29	**63.86**	33.87	73.6	68.85	62.22	62.91	**79.04**	**80.08**	71.31
Cos	97.17	**93.69**	61.99	**52.11**	76.9	61.25	60.65	63.68	78.97	78.49	72.49
Can	96.40	90.44	63.76	35.04	77.74	71.55	66.27	**70.61**	79.02	74.46	**72.53**
Che	**97.33**	91.69	60.59	32.59	72.74	68.74	61.92	62.23	78.95	79.51	70.63
Avg	**96.96**	91.21	**62.72**	36.83	**75.53**	68.57	63.24	**65.49**	**79.00**	77.96	

b. *Perbandingan Algoritma*

Tabel 2 Perbandingan Algoritma

Metode	k-NN	LVQ
Euclidean	74.89	67.76
Manhattan	**76.76**	67.7
Minkowski	75.21	67.4
Cosine	75.14	**69.84**
Canberra	76.64	68.42
Chebishev	74.31	66.95

c. Grafik Perbandingan pada Algoritma



Gambar 6 Grafik Perbandingan

4.2. Pembahasan

Berdasarkan Gambar 5.1 dapat diketahui bahwa pada algoritma k-NN metode perhitungan jarak yang memiliki

akurasi rata-rata tertinggi adalah Manhattan *Distance* yaitu sebesar 76.76%. Sedangkan pada algoritma LVQ metode perhitungan jarak yang memiliki akurasi rata-rata tertinggi adalah *Cosine Distance* yaitu sebesar 69.84%.

5. PENUTUP

5.1. Kesimpulan

Kesimpulan dari hasil penelitian yang telah dilakukan yaitu:

1. Pada perbandingan secara keseluruhan metode perhitungan jarak yang memiliki akurasi tertinggi adalah Canberra Distance yaitu sebesar 72.53%.
2. Pada algoritma k-*NN* metode perhitungan jarak yang memiliki akurasi rata-rata tertinggi adalah Manhattan Distance yaitu sebesar 76.76%.
3. Pada algoritma LVQ metode perhitungan jarak yang memiliki akurasi rata-rata tertinggi adalah Cosine Distance yaitu sebesar 69.84%.
4. Algoritma k-*NN* sangat cocok untuk dataset yang memiliki banyak kelas dan jumlah data yang sedikit sehingga mempercepat saat proses klasifikasi
5. Algoritma LVQ sangat cocok untuk dataset yang sangat besar, dengan proses pelatihan untuk mendapatkan bobot sehingga tidak mempengaruhi proses klasifikasi.

5.2. Saran

Adapun saran yang diberikan agar dapat dijadikan acuan penelitian selanjutnya yaitu:

1. Pengujian lebih mendalam terhadap pengaruh multiclass atau binaryclass pada dataset yang digunakan terutama data besar dengan jumlah data yang tidak seimbang.
2. Penelitian lebih lanjut menggunakan metode evaluasi lainnya seperti Leave-OneOut Cross Validation

DAFTAR PUSTAKA

[1] Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An Improved Method to construct basic probability assignment based on the confusion matrix for classification problem. Information Sciences 340-341.

[2] Faisal, M. R. & Nugrahadi, D. T. (2019), Belajar Data Science, Klasifikasi Dengan Bahasa R, Kalimantan Selatan : Scripta Cendekia.

[3] Foreman, J. (2014), Cosine Distance Cosine Similarity Angular Cosine Distance Angular Cosine Similarity, https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/cosdist.htm (diakses tanggal 29 Oktober 2019).

[4] Kusrini & Luthfi, E. T. (2009), Algoritma Data Mining, Yogyakarta: Andi Publishing.

[5] Kusumadewi, S. (2003), Artificial Intelligence (Teknik dan Aplikasinya). Yogyakarta: Graha Ilmu.

[6] Lazwardi, R. T. (2018), 4 Cara Menghitung Jarak dan Algoritma K-NN, https://belajarkalkulus.com/clustering-part-iii/ (diakses tanggal 29 Oktober 2019).

[7] Nurkhozin, A., dkk. (2011), Komparasi Hasil Klasifikasi Penyakit Diabetes Mellitus Menggunakan Jaringan Syaraf Tiruan Backpropagation dan 104 Learning Vector Quantization. Prosiding Seminar Nasional Penelitian, Pendidikan dan Penerapan FMIPA UNY. 14 Mei 2011.

[8] Pulungan, A. F. (2019), Analisis Kinerja Bray Curtis Distance Canberra Distance Dan Euclidean Distance Pada Algoritma K-Nearest Neighbor, Medan: Tesis, Rogram Studi S2 Teknik Informatika Fakultas Ilmu Komputer Dan Teknologi Informasi Universitas Sumatera Utara.

[9] Vashistha, R., & Nagar, S. (2017). An intelligent system for clustering using hybridization of distance function in learning vector quantization algorithm. 2017 Second International Conference onElectrical,Computerand Communication Technologies (ICECCT), pp. 1-7.

[10] Widiarsana, O., Putra, N.W., Budiyasa, P.G.I. dan Bismantara, A.N.I., Mahajaya, S.N. (2011), Data Mining: Metode Clasification K-Nearest Neighbor (K-NN).