PERBANDINGAN PENGARUH PENGGUNAAN EUCLIDIAN, MANHATTAN, DAN CHEBYCHEV TERHADAP TINGKAT AKURASI KLASIFIKASI

Puspa Miladin Nuraida Safitri A. Basid[1], Nadia Roosmalita Sari[2]

[1,2] Fakultas Teknologi Informasi, Universitas Merdeka Malang
Email: puspamiladin2808@gmail.com[1], nadiaroosmalitasari@gmail.com[2]

Abstrak

Kebutuhan sebuah sistem untuk klasifikasi sangat dibutuhkan untuk membantu mempermudah pekerjaan manusia. Klasifikasi adalah meramalkan sebuah objek dari klass yang belum memiliki kelas. Banyak ilmuan yang mulai memperkenalkan algoritma untuk klasifikasi, seperti Artificial Neural Network, Support Vector Machine, K-Nearest Neighbor, dan lain sebagainya. Metode ini sangatlah sederhana, dengan berdasarkan jarak terdekat antar data pembelajaran dan objek. Jarak ini lah yang digunakan sebagai nilai kemiripan atau kedekatan antara data uji. Permasalahan akan muncul ketika kita salah memilih metric jarak pada sebuah kasus yang akan kita klasifikasikan. Akibatnya adalah penurunan tingkat akurasi dari klasifikasi yang kita lakukan. Dalam penelitian ini dapat disimpulkan bahwa pada ada banyak alternatif metric jarak yang dapat di gunakan dalam klasifikasi menggunakan algoritma *K-Nearest Neighbor*. Perubahan penggunaan metric jarak standart memberikan hasil perubahan akurasi yang cukup signifikan. Dari percobaan di atas juga penulis dapat menyimpulkan bahwa metric jarak Manhattan memiliki akurasi yang lebih baik di banding metric jarak lainya pada studi kasus *Winconsin Breast Cancer*.

Kata kunci : klasifikasi, *K-Nearest Neighbor*, metric jarak, Manhattan.

Abstract (Example)

The necessary for a classification system is needed to help facilitate human task. Classification is to predict an object from a class that has no class. Many scientists are beginning to introduce algorithms for classification, such as Artificial Neural Network, Support Vector Machine, K-Nearest Neighbor,etc. This method is very simple, based on the closest distance between learning data and objects. This distance is used as the value of similarity or closeness between the testing data. Problems will appear we wrong to choose the metric of distance in a case that we will classify.. The result is a decrease in the accuracy of our classification. In this study it can be concluded that there are many alternative distance metrics that can be used in classification using K-Nearest Neighbor algorithm. Changes in the use of standard distance metrics provide significant changes in accuracy. From the above experiments also the authors can conclude that Manhattan's distance metrics have better accuracy compared to other distance metrics in the case studies of Breast Cancer.

Keyword: *classification, K-Nearest Neighbor, metric of distance, Manhattan*

PENDAHULUAN

Pada saat ini kebutuhan sebuah sistem untuk klasifikasi sangat dibutuhkan untuk membantu mempermudah pekerjaan manusia. Umumnya cara kerja dari sebuah sistem klasifikasi ini dikatakan cukup membantu karena dapat menggantikan peran pakar. Dimana proses klasifikasi yang

 ISSN : 2597 - 4696

dilakukan pakar secara subjektif di anggap cukup memakan waktu yang cukup lama. Menurut Han [1], klasifikasi adalah meramalkan sebuah objek dari klass yang belum memiliki kelas. Langkah awal yang dilakukan pada teknik klasifikasi adalah menentukan fitur-fitur. Fitur yang digunakan adalah fitur-fitur terbaik yang cukup menggambarkan perbedaan dengan ruang besar dari setiap fitur. Banyak ilmuan yang mulai memperkenalkan algoritma untuk klasifikasi, seperti Artificial Neural Network, Support Vector Machine, K-Nearest Neighbor, dan lain sebagainya. Dari sekian banyak algoritma klasifikasi yang ada, K-Nearest Neighbor merupakan salah satu algoritma yang cukup populer digunakan sebagai metode klasifikasi.

Metode ini di anggap cukup mudah dan efektif. Metode ini merupakan metode yang menggunakan data pembelajaran untuk menentukan kelas dari objek yang ingin diketauhi kelasnya. Metode ini sangatlah sederhana, dengan berdasarkan jarak terdekat antar data pembelajaran dan objek. Jarak ini lah yang digunakan sebagai nilai kemiripan atau kedekatan antara data uji. Pada kenyataanya, seringkali penggunakaan metode K-Nearest Neighbor sering di iringi dengan penggunaan metric jarak Euclidean [2]. Padahal tidak semua data pembelajaran memiliki akurasi klasifikasi yang baik ketika menggunakan metric ini. Permasalahan akan muncul ketika kita salah memilih metric jarak pada sebuah kasus yang akan kita klasifikasikan. Ditambah lagi dengan tidak sesuainya fitur dalam data pembelajaran yang kita gunakan. Akibatnya adalah penurunan tingkat akurasi dari klasifikasi yang kita lakukan. Oleh karena itu kita perlu belajar untuk memilih metric jarak terbaik untuk studi kasus atau data pembelajaran yang akan kita gunakan.

Alternatif metric jarak sendiri sudah banyak di temukan, tidak terbatas hanya dengan metric jarak euclidian. Sehingga kita diharapkan dapat memilih metric jarak dengan hasil akurasi terbaik untuk data pembelajaran yang kita pilih. Dalam paper ini, penulis mencoba untuk menentukan metric dengan hasil akurasi klasifikasi terbaik. Penulis menggunakan beberapa jenis data pembelajaran namun masih dalah satu lingkup kategori. Data pembelajaran yang akan di gunakan adalah data pembelajaran dari UCI.

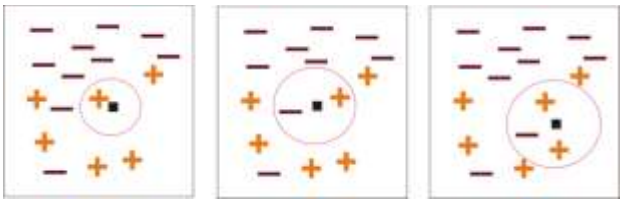
KAJIAN LITERATUR

K-Nearest Neighbor

Algortima *K-Nearest Neighbor* merupakan algoritma yang melakukan klasifikasi berdasarkan jarak suatu data

ISSN : 2597 - 4696

dengan dengan data yang lain. Dilihat dari cara kerjanya, algortima ini bekerja berdasarkan jarak terpendek dari data yang belum di ketahui kelasnya dengan data pembelajaran. Pada algortima *K-Nearest Neighbor*, selain metric jarak kita juga perlu menentukan nilai K. Nilai K sendiri merupakan K-data terdekat dari data pembelajaran.



Gambar 1.K-Nearest Neighbor dengan nilai K-tetangga: (a) 1-NN, (b) 2-NN, dan (c) 3-NN

Secara singkat langkah-langkah yang digunakan dalam metode *K-Nearest Neighbor* adalah menentukan nilai K. Masalah berikutnya adalah bagaimana menentukan nilai K yang tepat. Dalam paper ini penulis menentukan nilai K dengan menggunakan *role of thumb.* Yaitu nilai K sama dengan akar kuadrat dari jumlah sampel dari himpunan data pembelajaran[3].

$$k \approx \sqrt{\frac{n}{2}} \quad (1)$$

Dimana n adalah jumlah sample dari himpunan data pembelajaran.

Setelah nilai K ditentukan, selanjutnya adalah menghitung jarak dari masing-masing objek terhadap data pembelajaran yang sudah ada. Pada percobaan ini kali ini data pembelajaran yang di gunakan adalah Wisconsin Diagnostic *Breast cancer* (WDBC) dan Wisconsin Prognostic *Breast cancer* (WPBC) dari UCI. Sedangkan untuk penggunakan jarak akan di jelaskan pada sub bab selanjutnya. Langkah selanjutnya adalah mengurutkan objek-objek tersebut dalam kelompok yang memiliki jarak terkecil. Kemudian langkah terakhir adalah mengumpulkan kedalam kategori yang mayoritas atau paling banyak keluar.

Pengukuran Jarak

Setelah K ditentukan, berikutnya adalah mencari kuadrat jarak dari masing-masing objek terhadap data pembelajaran yang sudah ada. Dalam paper ini penulis menggunakan 3 buah metode pengukuran jarak, yaitu Euclidian, Manhattan, dan Chebychev Distance Metric.

Euclidian:

$$dist(p,q) = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2} \quad (2)$$

Dimana *n* adalah jumlah *attribute* dan p_k dan q_k adalah atribut ke-*k*

ISSN : 2597 - 4696

Manhattan:

$$dist(p,q) = \sum_{k=1}^{n} |p_k - q_k| \quad (3)$$

Chebychev:

$$d(p,q) = \max_{k=1}^{N} \left(|p_k - q_k| \right) \quad (4)$$

Perhitungan Akurasi

Pada penelitian ini, penghitungan akurasi di tentukan dengan membandingkan data yang diklasifikasikan dengan benar (*true positive*) dengan data pembelajaran. Sesuai persamaan di bawah ini:

$$accu = \frac{Correctly\ Classified\ Instances}{Total\ Number\ of\ Instances} \quad (5) \quad \%$$

IMPLEMENTASI

Uji Coba

Pada percobaan kali ini, penulis menggunakan data pembelajaran dari UCI dengan studi kasus Winconsin *breast cancer*. Dimana akan ada 2 jenis data pembelajaran yaitu data pembelajaran tentang diagnostic *breast cancer* dan data pembelajaran tentang prognostic *breast cancer*. Sebagai pengetahuan umum saja, diagnostic dapat di sebut sebagai deteksi awal apakah pasien tersebut memiliki ciri-ciri penderita *breast cancer*. Setelah diagnostic dilakukan, tahap selanjutnya adalah prognostic dimana tahap ini adalah menentukan langkah penanganan selanjutnya dari ciri-ciri yang di alami penderita *breast cancer* ini. Mengapa penulis memilih 2 jenis data pembelajaran yang saling berkaitan, nantinya diharapkan dapat memberikan solusi alternative pemilihan metric jarak untuk klasifikasi dengan *K-Nearest Neighbor* pada bidang *breast cancer* dengan 2 jenis data pembelajaran namun masih dalam satu kesatuan bidang.

Percobaan pertama dilakukan dengan data pembelajaran diagnostic *breast cancer*. Dimana jumlah data pembelajaran adalah sebanyak 567 dengan 30 jenis fitur. Kemudian K ditentukan tiap juamlah data uji yang diggunakan.

K	Skenario Uji	Akurasi		
		Euclidian	Manhattan	Chebychev
17	Cross Validation Fold-10	96.188 %	97.1336 %	93.6731 %

Table 1. Tabel Percobaan pada data uji Diagnostic *Breast Cancer* dengan 3 jenis metric jarak

Percobaan kedua dilakukan dengan data pembelajaran prognostic *breast cancer*. Dimana jumlah data pembelajaran sebanyak 198 dengan 32 jenis fitur.

ISSN : 2597 - 4696

K	Skenario Uji	Akurasi		
		Euclidian	Manhattan	Chebychev
10	Cross Validation Fold-10	75.7576 %	77.2727 %	74.2424 %

Table 2. Tabel Percobaan pada data uji Prognostic *Breast Cancer* dengan 3 jenis metric jarak

Dari percobaan yang sudah di lakukan, penulis mendapatkan nilai prosentase akurasi dari Diagnostic *Breast Cancer* yang terbaik adalah sebesar 97.1336 %. Nilai tersebut di dapatkan dengan menggunakan metric distance Manhattan. Pada percobaan kedua, yaitu pada Prognostic *Breast Cancer* di dapatkan nilai akurasi terbaik sebesar 77.2727 %. Nilai prosentase tersebut diperoleh dengan menggunakan metric distance Manhattan.

KESIMPULAN

Setelah melakukan percobaan pada 2 jenis data pembelajaran yang berbeda. Penulis menyimpulkan bahwa pada ada banyak alternatif metric jarak yang dapat di gunakan dalam klasifikasi menggunakan algoritma *K-Nearest Neighbor*. Perubahan penggunaan metric jarak standart memberikan hasil perubahan akurasi yang cukup signifikan. Dari percobaan di atas juga penulis dapat menyimpulkan bahwa metric jarak Manhattan memiliki akurasi yang lebih baik di banding metric jarak lainya pada studi kasus Winconsin *Breast Cancer*, baik diagnosis maupun prognosis. Penulis menganggap masih perlu dilakukan penelitian lagi tentang bagaimana mendapatkan akurasi terbaik dari klasifikasi menggunakan algoritma *K-Nearest Neighbor*. Tidak hanya dilihat dari metric jarak yang di gunakan, tetapi juga dari data uji yang akan digunakan. Sehingga perlu penelitian lebih lanjut tentang hubungan antara data uji, metric jarak dan akurasi pada algoritma algoritma *K-Nearest Neighbor*.

DAFTAR PUSTAKA

F. Gu, O. Liu, and X. Wang, "Semi-supervised weighted distance metric learning for kNN classification," *2010 Int. Conf. Comput. Mechatronics, Control Electron. Eng.*, pp. 406–409, 2010.

J. Han and K. Michelline, *Data Mining: Concepts And Techniques*, 2nd ed. San Francisco: Elsevier Inc., 2006.

M. Jirina and M. Jirina.Jr, "Classifiers Based on Inverted Distances," in *New Fundamental Technologies in Data Mining*, K. Funatsu, Ed. InTech, 2011, p. 369

ISSN : 2597 - 4696