ISSN: 1907-7912 EISSN: 2622-8092

Fuzzifikasi Data untuk Normalisasi Atribut dalam Perhitungan Algoritma K-Nearest Neighbour

M. Adib Al Karomi*
STMIK Widya Pratama Pekalongan
E-mail: adib.comp@gmail.com

RINGKASAN

K-Nearest Neighbour Merupakan algoritma yang sering digunakan dalam proses klasifikasi. Dalam proses perhitungannya algoritma ini menggunakan pendekatan similarity antar record atribut. Fungsi ini terbukti baik digunakan dan dapat menghasilkan klasifikasi yang cukup akurat. Kelemahan pendekatan similarity ini adalah apabila terdapat atribut dengan range nilai yang berbeda jauh maka akan menghasilkan nilai similarity yang besar. Nilai ini jelas tidak adil apabila terdapat atribut lain yang memiliki range sangat kecil. Perhitungan menggunakan fuzzy dinilai sangat cocok untuk menangani masalah ini. Dalam perhitungan fuzzy digunakan nilai terbesar yaitu 1 dengan nilai terendah adalah 0. Penelitian ini melakukan perhitungan algoritma K-Nearest Neighbour menggunakan fuzzy dan dilakukan perbandingan dengan perhitungan tanpa menggunakan fuzzifikasi data. Hasil dari penelitian ini membuktikan bahwa fuzzifikasi data untuk normalisasi atribut dapat membuat perhitungan klasifikasi k-nearest neighbor lebih akurat dan sesuai dengan sasaran.

Kata Kunci : KNN, Similarity, Normalisasi atribut

1. PENDAHULUAN

1.1 Latar Belakang

Big Data merupakan tema penelitian yang marak pada revolusi industry 4.0 ini. Salah satu proses yang tidak lepas dari big data adalah proses klasifikasi. Proses seperti ini dapat memungkinkan data yang sebelumnya tidak berguna menjadi lebih bernilai (Prasetyo, 2012). Banyak metode atau algoritma yang dapat digunakan dalam proses klasifikasi salah satunya adalah K-Nearest Neighbor. Algoritma K-Nearest Neighbor termasuk dalam salah satu algoritma klasifikasi terbaik serta mudah digunakan (Wu, 2009). Proses perhitungan dalam penggunaan algoritma ini dapat dikatakan sangat mudah dipahami dan untuk diimplementasikan. Secara garis besar proses perhitungan menggunakan algoritma ini adalah menghitung kedekatan atau similarity untuk semua atribut data training dengan atribut data testing yang ditanyakan. Nilai kedekatan untuk semua record yang ada dicatat lalu diambil nilai similarity terkecil dari semua record tersebut. Record dengan nilai similarity terkecil nantinya akan digunakan sebagai hasil klasifikasi. Nilai k dalam K-Nearest neighbor adalah banyaknya jumlah record terdekat yang akan diambil dan digunakan sebagai hasil klasifikasi. Apabila nilai k>1 maka hasil klasifikasi diputuskan dengan menggunakan hasil terbanyak dalam proporsi k yang ada.

Beberapa aplikasi dengan menggunakan perhitungan dasar K-Nearest neighbor banyak dijumpai di internet. Dalam perhitungannya banyak diantara penulis menggunakan fungsi similarity dan menggunakan nilai k sebesar 1 (k=1). Perhitungan untuk klasifikasi calon mahasiswa pernah dilakukan dan menggunakan seleksi fitur dikarenakan banyaknya atribut yang ada sehingga dapat mempersulit proses perhitungan (Alkaromi, 2014). Proses ini dilakukan karena banyak atribut data yang dinilai tidak layak untuk diikutsertakan dalam proses klasifikasi.

Salah satu permasalahan utama yang banyak dialami peneliti dalam penggunaan algoritma k-nearest neighbor adalah apabila terdapat beberapa atribut data numeric dengan range varian yang jauh antara satu atribut dengan atribut yang lain. Hal ini akan membuat nilai

similarity atau nilai kedekatan menjadi timpang antara satu atribut drngan atribut yang lain. Sebagai contoh atribut usia dengan tipe atribut numeric hanya memiliki range kedekatan maksimal 100tahun. Sedangkan atribut pendapatan dengan tipe yang sama dapat memiliki range kedekatan hingga jutaan bahkan miliyaran rupiah. Hal ini akan menimbulkan ketimpangan seandainya perbedaan usia 90 tahun yang memiliki nilai similarity 90 dibandingkan dengan perbedaan nilai pendapatan yang ratusan ribu rupiah. Sebenarnya perbedaan nilai pendapatan yang ratusan ribu tersebut tidak seberapa jika dibandingkan range pendapatan yang sangat tinggi.

Metode Fuzzy pertama kali ditemukan oleh Prof. Lotfi A. Zadeh dan banyak dikembangkan oleh peneliti karena kaunikan dan keunggulan metode tersebut. Sampai akhir hayatnya Zadeh terus mengembangkan keilmuannya dalam bidang matematis dan computer. Beberapa pengembangan dilakukan dengan menggunakan aplikasi fuzzy logic (Singh et al., 2013). Beberapa pendekatan fuzzy logic juga pernah dilakukan untuk objek penelitian yang lain termasuk system cerdas (Sets, 1997). Penelitian ini akan menggunakan penalaran fuzzy untuk melakukan normalisasi atribut dengan range yang berbeda.

2. METODE PENELITIAN

Penelitian ini menggunakan metode penelitian eksperimental. Dalam metode ini akan dilakukan perbandingan perhitungan K-Nearest neighbor dengan menggunakan similarity dan fuzzifikasi atribut dengan tipe numeric. Tahapan penelitian adalah sebagai berikut:

2.1 Analisa Atribut Data

Tahapan pertama yang dilakukan adalah melakukan analisa terhadap semua atribut data yang akan digunakan dalam proses klasifikasi. Proses analisa ini termasuk didalamnya adalah memilah atribut dengan tipe numeric untuk nantinya dilakukan perhitungan fuzzifikasi. Sedangkan untuk atribut dengan tipe nominal dilakukan perhitungan similarity secara manual.

2.2 Fuzzifikasi

Proses fuzzifikasi dilakukan dengan melalui beberapa tahapan. Tahapan pertama adalah melakukan perhitungan terhadap atribut numeric yang ada. Perhitungan ini meliputi pencatatan terhadap semua atribut numeric lalu membuat range untuk semua atribut. Selanjutnya nilai similarity antara record data testing dan data training dibagi dengan nilai range untuk tiap atribut. Proses ini akan membuat nilai similarity hanya memiliki range antara 0 sampai dengan 1. Similarity 0 akan muncul apabila data training sama dengan data testing. Sedangkan nilai similarity 1 akan muncul apabila jarak antara data training dan data testing yang muncul adalah jarak terjauh.

2.3 Perhitungan K-Nearest Neighbor

Proses berikutnya adalah perhitungan dengan menggunakan algoritma K-Nearest neighbor. Proses ini sebenarnya adalah melakukan perhitungan rata-rata untuk semua similarity dari atribut yang ada. Artinya apabila dalam klasifikasi terdapat 6 atribut maka nilai similarity dari keseluruhan atribut tersebut ditambahkan lalu dilakukan pembagian dari jumlah atribut yaitu 6.

3. HASIL DAN PEMBAHASAN

Objek data yang digunakan dalam penelitian ini adalah dataset play golf. Dataset ini memiliki 14 record dan 5 atribut dengan satu diantaranya adalah atribut label atau atribut tujuan klasifikasi. Atribut pertama adalah outlook dengan tipe nominal dan varian diantaranya ada sunny, overcast serta rain. Berikutnya atribut temperature dan humidity dengan tipe numeric yang memiliki range yang berbeda. Lalu atribut wind yang memiliki varian true dan false. Atribut terakhir yaitu atribut label (play) dengan varian atribut yaitu yes dan no. Tabel 1 merupakan dataset play golf.

Tabel 1. Dataset Play Golf

Outlook	Temperature	Humidity	Windy	Play
overcast	83	86	FALSE	yes
overcast	64	65	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
rainy	75	80	FALSE	yes
rainy	71	91	TRUE	no
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
sunny	75	70	TRUE	yes

Tabel 1 diatas selanjutnya akan digunakan sebagai data training. Sedangkan data testing yang ada adalah outlook=sunny, temperature=82, humidity=76, dan windy=true. Selanjutnya data testing tersebut akan dibandingkan serta digitung nilai similarity nya dengan semua data training pada tabel 1.

3.1 Hasil analisa atribut dataset

Dari dataset yang diperoleh terdapat 4 atribut regular yang yang nantinya dapat digunakan dalam proses klasifikasi. Atribut tersebut adalah outlook, temperature, humidity serta windy. Atribut outlook dan windy merupakan atribut dengan jenis nominal. Sedangkan atribut temperature dan humidity merupakan atribut dengan tipe numeric.

Perhitungan similarity tribut dengan tipe nominal dapat dilakukan secara manual karena tidak sesuai dengan karakteristik K-Nearest Neighbor (Witten, Frank, & Hall, 2011). Dalam hal ini atribut windy hanya memiliki 2 varian yaitu true dan false. Similarity atribut windy dapat diisikan 0 apabila isiannya sama, dan diisikan 1 apabila isiannya berbeda. Sedangkan untuk atribut outlook memiliki 3 varian yaitu overcast, sunny serta rainy. Sehingga harus diberikan nilai yang jelas. Dalam perhitungan penelitian ini diberikan nilai sebagaimana table 2 berikut.

Tabel 2. Nilai similarity atribut outlook

	Sunny	Overcast	Rainy
Sunny	**0**	0,6	1
Overcast	0,6	**0**	0,5
Rainy	1	0,5	**0**

Sedangkan untuk menghitung similarity atribut temperature dan humidity yang memiliki tipe numeric dilakukan perhitungan selisih antara data training dengan data testing. Dari data testing yang ada untuk record pertama dapat dihitung nilai similarity temperature |82-83| =1, Sedangkan untuk atribut humidity |86-76| =10. Nilai similarity pada record kedua untuk temperature |64-82| =18, dan untuk humidity |65-76| =11. Dan seterusnya untuk record berikutnya.

Hasil nilai similarity tersebut Nampak berfariasi dengan batasan maksimal yang tidak tentu. Hal ini jelas membuat nilai similarity menjadi tidak terstandarisasi dan memungkinkan terjadinya perhitungan yang melebihi batas normal.

3.2 Hasil Fuzzifikasi

Untuk proses fuzzifikasi sebenarnya dapat dilakukan dengan menghitung range atribut dengan tipe numeric. Nilai range tersebut digunakan untuk pembagi similarity antara dara testing dengan data training. Agar lebih mudah dapat digunakan pengurangan antara nilai tertinggi dengan nilai terendah dalam atribut tersebut. Untuk atribut temperature dapat dihitung nilai range adalah 85-65=20. Sedangkan untuk atribut humidity 96-65=31. Dari hal tersebut dapat dihitung nilai similarity dari record pertama atribut temperature adalah |82-83| /20 =**0,05** dan untuk atribut humidity |86-76| /31 =**0,3226**. Sedangkan untuk record yang kedua atribut temperature dapat dihitung dengan |64-82| /20 =**0,9** dan untuk atribut humidity dapat dihitung dengan |65-76| /31 =**0,3548**. Tabel 3 merupakan nilai similarity dan fuzzifikasi dari atribut temperature dan humidity terhadap data testing yang ada.

Tabel 3. Similarity dan fuzzifikasi atribut

Temperature	Sim.	Sim. Fuzzifikasi	Humidity	Sim.	Sim. Fuzzifikasi
83	1	0.05	86	10	0.32258065
64	18	0.9	65	11	0.35483871
72	10	0.5	90	14	0.4516129
81	1	0.05	75	1	0.03225806
70	12	0.6	96	20	0.64516129
68	14	0.7	80	4	0.12903226
65	17	0.85	70	6	0.19354839
75	7	0.35	80	4	0.12903226
71	11	0.55	91	15	0.48387097
85	3	0.15	85	9	0.29032258
80	2	0.1	90	14	0.4516129
72	10	0.5	95	19	0.61290323
69	13	0.65	70	6	0.19354839
75	7	0.35	70	6	0.19354839

3.3 Perhitungan K-Nearest Neighbor

Perhitungan menggunakan K-Nearest Neighbor dilakukan dengan bantuan Microsoft excel dan didapatkan hasil sebagaimana berikut. Tabel 4 merupakan hasil perhitungan similarity tanpa menggunakan fuzzifikasi. Sedangkan table 5 merupakan perhitungan similarity dengan fuzzifikasi atribut temperature dan humidity.

Tabel 4. Similarity seluruh record

Outlook	Sim.	Temperature	Sim.	Humidity	Sim.	Windy	Sim.	Play	Total Sim.
overcast	0.6	83	1	86	10	FALSE	1	yes	3.15
overcast	0.6	64	18	65	11	TRUE	0	yes	7.4
overcast	0.6	72	10	90	14	TRUE	0	yes	6.15
overcast	0.6	81	1	75	1	FALSE	1	yes	0.9
rainy	1	70	12	96	20	FALSE	1	yes	8.5
rainy	1	68	14	80	4	FALSE	1	yes	5
rainy	1	65	17	70	6	TRUE	0	no	6
rainy	1	75	7	80	4	FALSE	1	yes	3.25
rainy	1	71	11	91	15	TRUE	0	no	6.75
sunny	0	85	3	85	9	FALSE	1	no	3.25
sunny	0	80	2	90	14	TRUE	0	no	4
sunny	0	72	10	95	19	FALSE	1	no	7.5
sunny	0	69	13	70	6	FALSE	1	yes	5
sunny	0	75	7	70	6	TRUE	0	yes	3.25

Hasil kesimpulan menyatakan bahwa record nomor 4 merupakan record dengan nilai similarity terdekat dengan data testing dengan nilai similarity 0,9. Artinya hasil klasifikasi dapat dinyatakan memiliki kesimpulan akhir adalah YES.

Tabel 5. Fuzzifikasi similarity seluruh record

Outlook	Sim.	Temperature	Sim. Fuzzifikasi	Humidity	Sim. Fuzzifikasi	Windy	Sim.	Play	Total Sim.
overcast	0.6	83	0.05	86	0.3226	FALSE	1	yes	0.4931
overcast	0.6	64	0.9	65	0.3548	TRUE	0	yes	0.4637
overcast	0.6	72	0.5	90	0.4516	TRUE	0	yes	0.3879
overcast	0.6	81	0.05	75	0.0323	FALSE	1	yes	0.4206
rainy	1	70	0.6	96	0.6452	FALSE	1	yes	0.8113
rainy	1	68	0.7	80	0.1290	FALSE	1	yes	0.7073
rainy	1	65	0.85	70	0.1935	TRUE	0	no	0.5109
rainy	1	75	0.35	80	0.1290	FALSE	1	yes	0.6198
rainy	1	71	0.55	91	0.4839	TRUE	0	no	0.5085
sunny	0	85	0.15	85	0.2903	FALSE	1	no	0.3601
sunny	0	80	0.1	90	0.4516	TRUE	0	no	0.1379
sunny	0	72	0.5	95	0.6129	FALSE	1	no	0.5282
sunny	0	69	0.65	70	0.1935	FALSE	1	yes	0.4609
sunny	0	75	0.35	70	0.1935	TRUE	0	yes	0.1359

Dari table 5 setelah dilakukan fuzzifikasi ternyata muncul perbedaan yaitu record dengan nilai similarity terkecil adalah record nomor 14 dengan nilai similarity totalnya sebesar 0,1359.

Tabel 5 juga menggambarkan nilai similarity yang lebih terstandarisasi yaitu antara 0 untuk nilai yang sama persis dan 1 untuk nilai terjauh.

4. SIMPULAN DAN SARAN

4.1 Simpulan

Penelitian ini menghasilkan kesimpulan sebagaimana berikut:

1. Fuzzifikasi data pada atribut K-Nearest Neighbor dapat menormalisasi hasil similarity tiap atribut yang ada.
2. Hasil perhitungan similarity K-Nearest Neighbor dengan menggunakan pendekatan fuzzy dapat membuat klasifikasi lebih normal dan terstandarisasi.

4.2 Saran

Perhitungan K-Nearest Neighbor dalam penelitian ini tidak menggunakan pembobotan. Similarity dari semua atribut dijumlahkan dan hanya dibagi dengan jumlah atribut yang digunakan. Penelitian berikutnya dapat digunakan proses pembobotan agar tiap atribut memiliki kepentingan yang tidak sama.

5. DAFTAR PUSTAKA

Alkaromi, M. A. (2014). Information Gain untuk Pemilihan Fitur pada Klasifikasi Heregistrasi Calon Mahasiswa dengan Menggunakan K-NN.

Kurniawan, M. F., & Ivandari. (2017). Komparasi Algoritma Data Mining untuk Klasifikasi Kanker Payudara. *IC Tech*, *I April 20*, 1–8.

Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi menggunakan Matlab*. Yogyakarta: Andi Offset.

Sets, F. (1997). Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, *90*, 111–127.

Singh, H., Gupta, M. M., Meitzler, T., Hou, Z., Garg, K. K., Solo, A. M. G., & Zadeh, L. A. (2013). Real-Life Applications of Fuzzy Logic, *2013*.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition* (Vol. 40). Elsevier. https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

Wu, X. (2009). *The Top Ten Algorithms in Data Mining*. (V. Kumar, Ed.). New York: Taylor & Francis Group, LLC.