Optimasi Algoritma Naïve Bayes dengan Information Gain Ratio untuk Menangani Dataset Berdimensi Tinggi

M. Adib Al Karomi, Ivandari
STMIK Widya Pratama Pekalongan
adib.comp@gmail.com, ivandarialkaromi@gmail.com

RINGKASAN

Perkembangan ilmu komputer sekarang memungkinkan adanya pencatatan semua proses bisnis di segala bidang dengan media penyimpanan yang besar. Data di bidang atronomi, kesehatan, ekonomi, pemerintahan dan sebagainya banyak tercatat dan semakin banyak dari tahun ke tahun. Data mining merupakan ilmu yang dapat mengolah data menjadi sebuah representasi pengetahuan dengan menggunakan beberapa metode atau algoritma matematis. Salah satu fungsi utama data mining adalah klasifikasi. Dalam proses klasifikasi semua data lama digunakan sebagai data pembelajaran untuk menyimpulkan data baru yang belum sepenuhnya diketahui. Data yang sebelumnya tidak memiliki makna dapat menjadi sebuah pengetahuan baru dengan menggunakan klasifikasi data mining. Banyak algoritma yang dapat digunakan dalam proses klasifikasi. Salah satu algoritma yang terbukti baik untuk proses klasifikasi data berdimensi tinggi adalah naïve bayes. Dalam data berdimensi tinggi banyaknya atribut data dapat mempengaruhi hasil klasifikasi. Banyaknya atribut data yang relevan dapat meningkatkan performa algoritma. Sedangkan banyaknya atribut data yang tidak relevan dapat menurunkan tingkat akurasi sebuah algoritma. Dari hasil penelitian ini diketahui bahwa seleksi fitur information gain dapat meningkatkan performa klasifikasi naive bayes.

Kata Kunci : peningkatan performa bayes, information gain ratio, data public

1. PENDAHULUAN

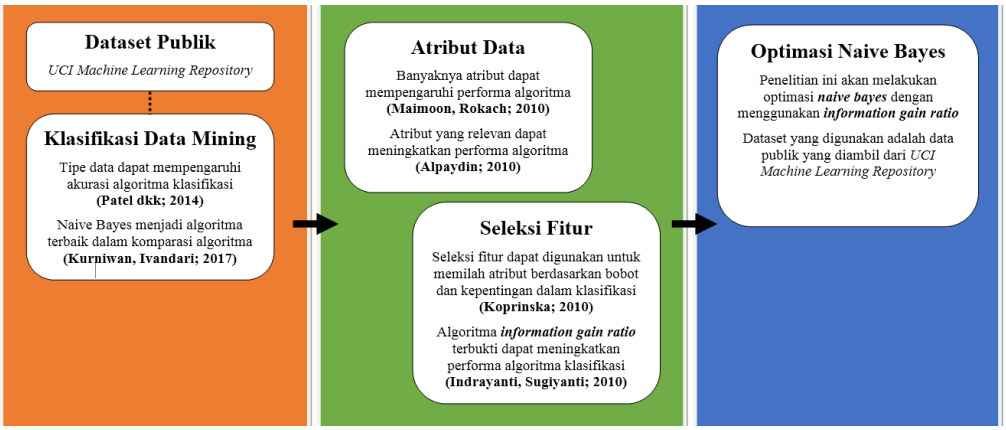
1.1 Latar Belakang Masalah

Data mining merupakan bidang ilmu yang mempelajari data lampau untuk diambil menjadi sebuah pengetahuan baru (Witten, Frank, and Hall 2011). Salah satu fungsi utama data mining adalah klasifikasi. Seiring perkembangan teknologi komunikasi, objek penelitian data mining menjadi sangat luas. Data mining banyak digunakan dalam proses klasifikasi modern dengan menggunakan kumpulan data yang ada. Tipe data dapat mempengaruhi performa suatu algoritma klasifikasi data mining (Amancio et al. 2013). Algoritma klasifikasi terbaik untuk sebuah data belum tentu baik apabila digunakan untuk mengolah data yang lain (Patel, Vala, and Pandya 2014). Perbedaan performa algoritma ini dikarenakan adanya karakteristik yang berbeda dalam sebuah data (Ragab et al. 2014) (Ashari, Paryudi, and Tjoa 2013).

Salah satu algoritma klasifikasi terbaik dan banyak digunakan peneliti adalah naive bayes (Wu 2009). Naive bayes terbukti dapat menangani atribut data nominal dengan memanfaatkan perhitungan probabilitasnya. Dalam penelitian lain yang melakukan komparasi algoritma klasifikasi naive bayes terbukti menjadi algoritma terbaik dengan tingkat akurasi tertinggi (Kurniawan and Ivandari 2017). Dalam sebuah proses klasifikasi dilakukan perhitungan untuk setiap atribut data yang ada. Perbedaan jumlah atribut data yang digunakan dapat mempengaruhi tingkat akurasi sebuah algoritma (Maimoon and Rokach 2010). Beberapa atribut yang relevan dan sesuai dapat meningkatkan performa suatu algoritma, sedangkan adanya atribut data yang tidak relevan dapat membuat performa algoritma menurun dan berkurangnya tingkat akurasi sebuah algoritma (Han and Kamber 2006). Tipe dari atribut dataset yang digunakan dalam proses klasifikasi juga dapat mempengaruhi tingkat akurasi sebuah algoritma (Alpaydin 2010).

Salah satu proses untuk memilih atribut yang akan digunakan dalam proses klasifikasi adalah dengan melakukan pre processing yaitu seleksi fitur. Seleksi fitur adalah perlakuan terhadap dataset untuk menghitung kepentingan seluruh atribut data yang ada. Seleksi fitur ini dapat menghasilkan tingkat kepentingan atau biasa disebut dengan bobot untuk semua atribut dalam dataset. Atribut dengan nilai bobot tinggi berarti memiliki kepentingan yang tinggi, begitu juga sebaliknya. information gain ratio merupakan metode seleksi fitur yang banyak digunakan dan terbukti dapat menangani dataset berdimensi tinggi (Koprinska 2010). Metode ini merupakan pembaruan dari metode lama yaitu information gain yang juga terbukti baik digunakan untuk menangani dataset berdimensi tinggi (Alkaromi 2014). Penelitian akan menggunakan information gain ratio untuk optimasi algoritma naive bayes dalam melakukan klasifikasi dataset berdimensi tinggi.

Dalam penelitian ini akan digunakan 10 macam dataset publik terpopuler yang biasa digunakan dalam penelitian klasifikasi data mining. Data publik yang akan digunakan diambil dari UCI Machine Learning Repository. UCI Repository merupakan penyedia dataset publik yang telah teruji dan banyak digunakan dalam penelitian tingkat internasional. Website dari penyedia dataset dapat diakses di: https://archive.ics.uci.edu/ml/index.php.



Gambar 1 Kerangka pemikiran penelitian

2. METODE PENELITIAN

Metode penelitian dalam penelitian ini adalah eksperimental. Dalam penelitian ini akan dilakukan pengukuran akurasi semua dataset yang ada menggunakan algoritma naïve bayes. Setelah diketahui performa algoritma dalam melakukan klasifikasi pada tiap dataset, berikutnya akan dilakukan optimasi menggunakan information gain ratio untuk meningkatkan akurasi algoritma serta memperbaiki performa dari naïve bayes. Gambar 1 diatas merupakan kerangka pemikiran dalam penelitian ini

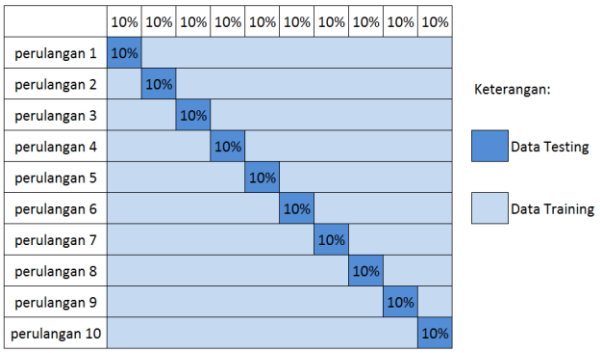
2.1 Tahapan Pengumpulan Data

Dalam tahapan ini akan dikumpulkan dataset yang merupakan data terpopuler yang digunakan dalam proses penelitian. Data yang digunakan dalam penelitian ini adalah dataset public dari UCI repository. UCI repository merupakan salah satu sumber dataset terpercaya yang menyediakan lebih dari 440 dataset machine learning. Dataset dari UCI banyak digunakan oleh peneliti bidang ilmu komputer untuk menguji metode atau model suatu algoritma. Dataset tersebut dapat diunduh pada laman: https://archive.ics.uci.edu/ml/datasets.html.

Selain pengumpulan dataset, dalam tahapan ini juga akan dilakukan analisa terhadap semua dataset yang telah diperoleh. Analisa tersebut terkait jenis atribut serta tipe dataset terpilih. Analisa ini juga didasarkan dari referensi terkait yaitu dari artikel ilmiah yang membahas tentang semua dataset terpilih.

2.2 Desain dan Pemodelan Algoritma

Dalam tahapan desain dan pemodelan algoritma ini akan dilakukan menggunakan pre processing seleksi fitur dilanjutkan dengan klasifikasi naive bayes. Dalam tahapan ini juga akan digunakan X-Validation untuk melakukan validasi dataset. Validasi dilakukan dengan cara membagi data menjadi 10 bagian, 9 diantaranya dijadikan data training dan 1 bagian yang lain digunakan sebagai data testing. Proses ini diulang sebanyak 10 kali sehingga semua bagian data pernah menjadi data testing. Proses ini biasa disebut dengan 10 folds cross validation. Gambar 2 adalah representasi dari 10 folds cross validation.



Gambar 2 Representasi 10 folds cross validation

2.3 Seleksi Fitur dan Perhitungan Akurasi Algoritma

Tahapan selanjutnya adalah seleksi fitur menggunakan information gain ratio. Dalam tahap ini akan dilakukan perhitungan bobot seluruh atribut data. Selanjutnya diterapkan treshold atau batas dari bobot yang akan digunakan. Atribut dengan bobot diatas treshold selanjutnya akan digunakan dalam proses klasifikasi. Sedangkan atribut yang memiliki bobot dibawak treshold nantinya tidak akan digunakan dalam proses klasifikasi. Aplikasi bantu dalam penelitian ini adalah rapid miner.

Proses perhitungan tingkat akurasi menggunakan confussion matrix atau matrix kebingungan. Dalam matrix ini nantinya akan muncul klasifikasi yang sesuai dengan data asli dan yang tidak sesuai dengan aslinya. Proses perhitungan akurasi didasarkan pada matrix ini. Semakin banyak klasifikasi yang sesuai dengan data asli maka akurasi semakin baik.

2.4 Pengujian

Dalam tahapan ini akan dicatat perubahan akurasi yang terjadi pada algoritma klasifikasi naive bayes. Perbandingan ini dilakukan antara hasil akurasi naive bayes menggunakan seluruh atribut dataset dengan hanya menggunakan atribut terpilih setelah dilakukan seleksi fitur information gain ratio. Pencatatan dilakukan terhadap seluruh dataset untuk mengetahui peningkatan performa naive bayes dengan melakukan seleksi fitur information gain ratio. Dalam tahapan pengujian ini akan diketahui optimasi algoritma naive bayes dengan menggunakan algoritma information gain ratio.

3. HASIL DAN PEMBAHASAN

3.1 Dataset dan Hasil Pembobotan

Penelitian ini menggunakan 10 dataset yang diambil dari data public. Keseluruhan dataset tersebut merupakan dataset yang sudah terbukti dan digunakan untuk pengujian metode oleh banyak peneliti dunia. Dataset yang didapat kemudian dilakukan pembobotan menggunakan information gain ratio sebagaimana berikut:

3.1.1 Bobot Atribut Breast Cancer Data Set

Tabel 1. Perhitungan IGR *breast cancer dataset*

ATTRIBUTE	WEIGHT
Menopaose	0
breast	0.008750065
Age	0.040894777
breast-quad	0.067550806
Tumor Size	0.203724543
irradiat	0.366824468
inv-nodes	0.600839672
node-caps	0.693480969
deg-mailing	1

3.1.2 Bobot Atribut Primary Tumor Data Set

Tabel 2. Perhitungan IGR *primary tumor dataset*

ATTRIBUTE	WEIGHT
age	0
supraclavicular	0.004891798
degree-of-diffe	0.022819926
bone-marrow	0.03172934
lung	0.033871912
pleura	0.033871912
skin	0.041285759
brain	0.043627757

bone	0.04777811
axillar	0.050046303
sex	0.075037593
mediastinum	0.090688323
peritoneum	0.092963833
liver	0.104237476
abdominal	0.109446099
histologic-type	0.280062103
neck	1

3.1.3 Bobot Atribut Car Evaluation Data Set

Tabel 3. Perhitungan IGR *car evaluation dataset*

ATTRIBUTE	WEIGHT
doors	0
lug_boot	0.014374799
maint	0.058699326
buying	0.070660691
persons	0.995349085
safety	1

3.1.4 Bobot Atribut Congressional Voting Records Data Set

Tabel 4. Perhitungan IGR *congressional voting records dataset*

ATTRIBUTE	WEIGHT
water-project-cost-sharing	0
immigration	0.00732874
export-administration-act-south-africa	0.11690109
synfuels-corporation-cutback	0.1381962
handicapped-infants	0.16713487
religious-groups-in-schools	0.20556517
anti-satellite-test-ban	0.25887584
duty-free-exports	0.26452891
superfund-right-to-sue	0.27480281
mx-missile	0.38124236
crime	0.43391963
education-spending	0.44329566
aid-to-nicaraguan-contras	0.44373255
el-salvador-aid	0.54352193
adoption-of-the-budget-resolution	0.58779323
physician-fee-freeze	1

3.1.5 Bobot Atribut Lymphography Data Set

Tabel 5. Perhitungan IGR *lyphography dataset*

ATTRIBUTE	WEIGHT
extravasates	0
bl. of lymph. c	0.02238301
dislocation of	0.041213236
exclusion of no	0.06123256
by pass	0.064357898
bl. of lymph. s	0.11902886
early uptake in	0.127117638
special forms	0.128804093
block of affere	0.150424799
lym.nodes enlar	0.151642277
no. of nodes in	0.239227512
changes in node	0.32562959
changes in stru	0.341382072
regeneration of	0.361617522
changes in lym	0.466187988
defect in node	0.621560606
lym.nodes dimin	0.8420194
lymphatics	1

3.1.6 Bobot Atribut Mushroom Data Set

Tabel 6. Perhitungan IGR *mushroom dataset*

ATTRIBUTE	WEIGHT
veil-type	0
stalk-surface-above-ring	0.002229477
stalk-shape	0.019496671
cap-color	0.036763127
cap-surface	0.046453373
cap-shape	0.075572019
stalk-color-below-ring	0.171266454
habitat	0.176490042
stalk-root	0.189318448
gill-attachment	0.209441182
ring-number	0.227708976
population	0.258052488
veil-color	0.318265366
gill-color	0.352226637
ring-type	0.366250147
stalk-color-above-ring	0.368463919
gill-spacing	0.404850909
stalk-surface-below-ring	0.497456509
bruises?	0.502857694
spore-print-color	0.558513265
gill-size	0.660303864
odor	1

3.1.7 Bobot Atribut Nursery Data Set

Tabel 7. Perhitungan IGR *nursey dataset*

ATTRIBUTE	WEIGHT
parents	0
has_nurs	0
form	0
children	0
finance	0
housing	1
social	1

3.1.8 Bobot Atribut SPECT Heart Data Set

Tabel 8. Perhitungan IGR *spect heart dataset*

ATTRIBUTE	WEIGHT
F9	0
F14	0.035981354
F19	0.164593427
F16	0.175421838
F3	0.177285507
F4	0.24776365
F8	0.265898841
F15	0.284647411
F5	0.346383075
F11	0.383383248
F1	0.4319535
F2	0.434825476
F6	0.434825476
F10	0.491758382
F12	0.515478315
F20	0.646766215
F18	0.664504878
F17	0.675712011
F21	0.812554429
F22	0.925480993
F7	0.985338618
F13	1

3.1.9 Bobot Atribut Tic-Tac-Toe Endgame Data Set

Tabel 9. Perhitungan IGR *tic-tac-toe endgame dataset*

ATTRIBUTE	WEIGHT
middle-left-square	0
middle-right-square	0
bottom-middle-square	0
top-middle-square	0.058200469
bottom-left-square	0.080180196
bottom-right-square	0.080180196
top-left-square	0.204078985
top-right-square	0.204078985
middle-middle-square	1

3.1.10 Bobot Atribut Trains Data Set

Tabel 10. Perhitungan IGR *trains dataset*

ATTRIBUTE	WEIGHT
num_loads	0
Rectangle_next_to_hexagon	0
Rectangle_next_to_circle	0
Hexagon_next_to_hexagon	0
Circle_next_to_circle	0
length	0.062947682
length 2	0.062947682
num_wheels 4	0.062947682
length 4	0.062947682
num_loads 4	0.062947682
Rectangle_next_to_rectangle	0.062947682
Triangle_next_to_circle	0.062947682
load_shape 2	0.133945891
load_shape 4	0.151830987
shape	0.274213038
num_wheels	0.366655922
num_wheels 2	0.366655922
num_loads 2	0.366655922
Triangle_next_to_hexagon	0.366655922
Hexagon_next_to_circle	0.366655922
load_shape	0.405297887
length 3	0.405297887
shape 4	0.42703593
Triangle_next_to_triangle	0.521347612
shape 2	0.544762684
num_wheels 3	0.551337676
shape 3	0.586566049
load_shape 3	0.600117372
num_loads 3	0.603503476
Number_of_cars	0.714908562
Number_of_different_loads	0.714908562
Rectangle_next_to_triangle	1

3.2 Perhitungan Tingkat Akurasi Naive Bayes

Proses perhitungan tingkat akurasi naïve bayes dilakukan dengan menggunakan aplikasi rapid miner. Dalam aplikasi ini semua dataset dikumpulkan dan dilakukan perhitungan. Perhitungan dilakukan dengan cara membandingkan hasil tingkat akurasi naïve bayes. Dataset yang digunakan adalah dataset original dan dataset yang telah dilakukan seleksi fitur untuk mengurangi atribut datanya. Dari hasil percobaan terhadap sepuluh dataset yang ada, didapatkan hasil sebagaimana tabel 11 berikut:

Tabel 11 Rekap hasil penelitian

No	Dataset	Tingkat akurasi (%)		Kenaikan tingkat akurasi
		Naïve bayes	Naïve bayes + IGR	
1	Breast cancer	70.64%	**73.07%**	2.43%
2	Primary tumor	94.09%	**97.09%**	3.00%
3	Car evaluation	**91.28%**	90.34%	-0.94%
4	Congressional voting	90.34%	**91.49%**	1.15%
5	Lymphography	74.38%	**76.29%**	1.91%
6	Mushroom agaricus-lepiota	**99.57%**	99.54%	-0.03%
7	Nursey	28.8%	**31.21%**	2.41%
8	Spect heart	**61.55%**	60.88%	-0.67%
9	Tic-tac-toe	69.83%	**73.07%**	3.24%
10	Trains-transformed	50%	**80%**	30.00%
				42.50%

Dari tabel 11 diatas dapat diketahui bahwa terdapat 3 dataset yang justru mengalami penurunan tingkat akurasi. Dataset car evaluation mengalami penurunan 0.94%, dataset mushroom agaricus-lepiota mengalami penurunan akurasi 0.03%, dan dataset spect heart mengalamia penurunan sebesar 0.67%. Ketiga dataset tersebut hanya mengalami penurunan akurasi kurang dari 1%. Terdapat 7 dataset yang mengalami kenaikan tingkat akurasi. Kenaikan akurasi dataset tersebut lebih dari 1% bahkan ada yang sampai naik 30%. Dari kesepuluh dataset public tersebut didapatkan rekapitulasi kenaikan akurasi sebesar 42.5% yang artinya rata-rata kenaikan tiap dataset dengan penggunaan information gain ratio adalah 4.25%.

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Telah diketahui tingkat kepentingan atau bobot masing masing atribut data dari semua dataset. Bobot tersebut didapatkan dari hasil perhitungan dengan menggunakan metode information gain ratio. Dari 10 dataset yang didapat, diketahui bahwa penggunaan information gain ratio dapat meningkatkan performa naïve bayes untuk dataset berdimensi tinggi. Rata peningkatan akurasi tiap dataset adalah 4.25%.

4.2 Saran

Dalam penelitian ini digunakan information gain ratio dengan penggunaan threshold secara manual. Berikutnya dapat dilakukan otomasi threshold untuk optimasi akurasi hasil klasifikasi.

5. DAFTAR PUSTAKA

Alkaromi, M Adib. 2014. "Information Gain Untuk Pemilihan Fitur Pada Klasifikasi Heregistrasi Calon Mahasiswa Dengan Menggunakan K-NN."

Alpaydin, Ethem. 2010. *Introduction to Machine Learning Second Edition*. London: The MIT Press.

Amancio, D. R., C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. a. Rodrigues, and L. Da F. Costa. 2013. "A Systematic Comparison of Supervised Classifiers," October. http://arxiv.org/abs/1311.0202v1.

Ashari, Ahmad, Iman Paryudi, and A Min Tjoa. 2013. "Performance Comparison between Naïve Bayes , Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool" 4 (11): 33–39.

Han, Jiawei, and Micheline Kamber. 2006. *Data Mining: Concepts and Techniques Second Edition*. *Elsevier*. Elsevier.

Koprinska, Irena. 2010. “Feature Selection for Brain-Computer Interfaces,” 100–111.

Kurniawan, M. Faisal, and Ivandari. 2017. “Komparasi Algoritma Data Mining Untuk Klasifikasi Kanker Payudara.” *IC Tech* I April 20: 1–8.

Maimoon, Oded, and Lior Rokach. 2010. *Data Mining and Knowledge Discovery Handbook*. Vol. 40. Springer. https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Patel, Kanu, Jay Vala, and Jaymit Pandya. 2014. “Comparison of Various Classification Algorithms on Iris Datasets Using WEKA” 1 (1): 1–7.

Ragab, Abdul Hamid M., Amin Y. Noaman, Abdullah S. Al-Ghamdi, and Ayman I. Madbouly. 2014. “A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining.” *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments - IDEE ’14*, 106–13. https://doi.org/10.1145/2643604.2643631.

Witten, Ian H, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition*. Elsevier.

Wu, Xindong. 2009. *The Top Ten Algorithms in Data Mining*. Edited by Vipin Kumar. New York: Taylor & Francis Group, LLC.