JURNAL DASI
Vol. 15 No. 1 Maret 2014
ISSN: 1411-3201

PERBANDINGAN KINERJA ALGORITMA KLASIFIKASI *NAÏVE BAYESIAN, LAZY-IBK, ZERO-R,* DAN *DECISION TREE- J48*

Sulidar Fitri

Teknik Informatika STMIK AMIKOM Yogyakarta
email : inboxfitri@gmail.com

Abstraksi

Penelitian ini difokuskan untuk mengetahui kinerja terbaik dari beberapa algoritma klasifikasi dalam data mining yaitu Naïve Bayesian, Lazy-IBK, Zero-R, dan *Decision Tree- J48. Aspek yang dilihat adalah dari sisi keakuratan prediksi dan kecepatan/efisiensi. Software yang digunakan untuk mengevaluasi beberapa algoritma klasifikasi tersebut adalah Weka versi 3.7.7. Hasil pengujian menunjukan bahwa algoritma naïve Bayesian memiliki akurasi terbaik sebesar 85,12% pada mode tes cross-validation. Namun algoritma ZeroR memiliki kecepatan terbaik untuk semua mode tes dan semua data set di dalam penelitian ini.*

***Kata Kunci* :**
Algoritma klasifikasi, Naïve Bayesian, Lazy-IBK, Zero-R, Decision Tree- J48

Pendahuluan

Pada era informasi dan era teknologi canggih seperti beberapa tahun terakhir ini dapat kita ketahui bahwa sudah banyak sistem terkomputerisasi yang dibangun dengan menggunakan desain database untuk data berskala besar. Apalagi didukung dengan tempat penyimpanan data yang sangat besar pula sehingga mendukung manusia untuk leluasa menyimpan semakin banyak data.

Semakin banyak data tersimpan maka bisa dikatakan telah terjadi penumpukan data yang sangat besar pada tempat penyimpanan. Kondisi data yang bertumpuk terus menerus ini akan sia-sia jika tidak dimanfaatkan kembali untuk kebutuhan informasi dimana kebutuhan informasi dari tahun ke tahun terus meningkat.

Untuk mengatasi tumpukan data berskala besar, digunakanlah suatu teknik penggalian informasi dari data yang sudah bertumpuk tersebut. Teknik yang disebut dengan *Data mining* bisa menjadi solusi untuk mengatasi tumpukan data yang ada pada tempat penyimpanan sehingga data-data tersebut dapat dimanfaatkan kembali tanpa terbuang percuma.

Data mining merupakan teknik yang sering digunakan untuk menggali informasi yang tersembunyi dalam data yang besar. Sehingga dengan menggunakan teknik data mining tersebut kita dapat menemukan informasi yang berupa pola, ciri, dan aturan atau dikenal sebagai istilah *knowledge*.

Pada proses *Data mining* terdapat beberapa metode pengolahan data, salah satunya adalah klasifikasi. Tujuan dari penelitian ini adalah untuk mengetahui perbandingan kinerja dari beberapa algoritma yang terdapat dalam metode klasifikasi sehingga dapat diketahui algoritma mana yang mempunyai keunggulan dalam hal keakuratan prediksi dan kecepatan/efisiensi.

Beberapa algoritma yang akan dibandingkan dalam penelitian ini adalah *Naïve Bayesian, Lazy-IBK, Zero-R,* dan *Decision Tree- J48.* Penelitian ini menggunakan *software* WEKA versi 3.7.7 sebagai alat bantu untuk mengevaluasi kinerja empat algoritma tersebtu.

Pada penelitian yang dilakukan oleh Youn dan McLeod di tahun 2006[4] membuktikan bahwa *decision tree* dengan algoritma C4.5 lebih efisien dan paling sederhana jika dibandingkan algoritma klasifikasi yang lainnya. Dari penelitian lain yang dilakukan oleh Jyh-Jian Sheu pada tahun 2008[3] diperoleh hasil bahwa metode ID3 dari *decision tree* merupakan metode yang paling baik jika dibandingkan dengan beberapa algoritma klasifikasi lainnya.

Dari kedua penelitian tersebut, dapat dilihat bahwa algoritma *decision tree* mempunyai kinerja yang unggul dibandingkan dengan algoritma klasifikasi yang lain, namun dalam penelitian ini akan membuktikan apakah hasil yang sama bisa didapatkan oleh algoritma *decision tree- J48*.

Tinjauan Pustaka

Data mining merupakan sebuah proses dari *knowledge discovery* (penemuan pengetahuan) dari

 ISSN: 1411-3201

data yang sangat besar [1]. Sementara itu Tan dkk. berpendapat bahwa data mining adalah proses secara otomatis untuk menemukan informasi yang berharga dari repositori data yang sangat besar [5]. Dengan demikian, dari tumpukan data tersebut akan didapat beragam informasi yang berharga dan penting yang sebelumnya tidak diketahui.

Ada beberapa teknik yang dimiliki data mining berdasarkan tugas yang bisa dilakukan, yaitu deskripsi, estimasi, prediksi, klasifikasi, klastering, dan asosiasi [2]. Namun penelitian ini hanya akan fokus pada metode klasifikasi. Klasifikasi merupakan teknik untuk mengelompokan data berdasarkan beberapa kategori tertentu. Pada metode klasifikasi juga terdapat beberapa algoritma diantaranya *Naïve Bayesian, Lazy-IBK, Zero-R*, dan *Decision Tree- J48* dan masih banyak algoritma lainnya namun tidak digunakan dalam penelitian ini.

Naive bayes classifier (NBC) merupakan salah satu metode pada teknik klasifikasi dan termasuk dalam classifier statistik yang dapat memprediksi probabilitas keanggotaan class. NBC berprinsip pada teori bayes. NBC mengasumsikan bahwa nilai atribut pada sebuah *class* adalah independen terhadap nilai pada atribut yang lain [1].

$$P(\mathrm{Ci} \mid X) = \frac{P(X \mid Ci)\,P(Ci)}{P(X)} \quad (1)$$

Class Ci adalah nilai terbesar, sedangkan P(X) adalah konstanta untuk semua *class.* P merupakan *posterior probability*.

Lazy-IBK atau dikenal dengan algoritma *K-NN(K-Nearest neighbor).* Algoritma *K-Nearest Neighbor* adalah pendekatan untuk mencari kasus dengan menghitung kedakatan antara kasus baru dengan kasus lama. Jumlah data/tetangga terdekat ditentukan oleh *user* yang dinyatakan dengan *k* [1].

Zero-R adalah algoritma untuk memprediksi kelas mayoritas data uji nilai(jika nominal) atau rata-rata (jika numerik) [1].

Decision tree adalah algoritma yang paling banyak digunakan untuk masalah klasifikasi. Sebuah *decision tree* terdiri dari beberapa simpul yaitu *tree's roo*t, *internal node* dan *leafs*. Konsep entropi digunakan untuk penentuan pada atribut mana sebuah pohon akan terbagi (*split*) [1]. Semakin tinggi entropy sebuah sampel, semakin tidak murni sampel tersebut. Rumus yang digunakan untuk menghitung entropy sampel S adalah

Entropy(S) = -p1log2p1–p2log2p2 (2)

Dimana : p1 adalah proporsi sampel atau grup 1 yang akan dipasangkan dengan proporsi p2 proporsi grup 2.

Metode Penelitian

Penelitian ini menggunakan 2 data set yang akan diklasifikasikan dalam bentuk format file .arff. Kedua data set tersebut diambil dari UCI *Data Repository* [6].

Dua data set tersebut adalah data set Ecoli dan data set Yeast. Kedua data set tersebut berisi tentang data lokalisasi protein pada bakteri EColi dan yeast (ragi). Detail keterangan dari masing-masing data set tertera pada tabel 1.

Hasil yang tertera pada jendela *classifier output* setelah melalui proses pembangunan model akan dicatat dan dari pencatatan tersebut akan dbandingkan nilainya, sehingga dapat diketahui algoritma mana yang kinerjanya paling baik.

Tabel 1. Informasi detail data set

Data Set	**EColi**	**Yeast**
Tipe File	ARFF	ARFF
Banyak Atribut	8	9
Banyak record	336	1484
Karakteristik Atribut	Real	Real
Karakteristik Data set	Multivariat	Multivariat
Missing Value	Tidak ada	Tidak ada

Pada data set EColi memiliki 336 baris data dan memiliki 8 atribut diantaranya: *SequenceName, mcg, gvh, lip chg, aac, alm1, alm2*. Pada data set Yeast memiliki 1448 baris data dan memiliki 9 atribut diantaranya: *SequenceName, mcg, gvh, alm,mit, erl, pox,vac, nuc*. Satu kolom kelas ada di kolom paling terakhri dari kedua data set tersebut.

Parameter yang digunakan untuk membandingkan kinerja dari beberapa algoritma klasifikasi adalah:

1) *Test Mode*: Mendefinisikan mode tes yang digunakan adalah *cross-validation test* dan *percentage split test mode* untuk teknik evaluasi.
2) *Time to build model*: merupakan istilah untuk menerangkan berapa waktu yang dibutuhkan untuk membangun model klasifikasi untuk masing-masing algoritma
3) *Correctly classified instances*: berapa banyak baris data yang terklasifikasikan dengan benar.
4) *Incorrectly classified instances*: berapa banyak baris data yang terklasifikasikan tidak benar.

 ISSN: 1411-3201

Hasil dan Pembahasan

Hasil evaluasi dari kinerja algoritma *Naïve Bayesian, Lazy-IBK, Zero-R,* dan *Decision Tree-J48* dapat dilihat pada tabel 2. Informasi yang didapat dari tabel 2 terdiri dari mode tes yang digunakan untuk masing-masing data set yang terdiri dari mode test *cross-validation* dan *percentage-split*. Menu mode tes yang digunakan adalah *default*.

Informasi ukuran akurasi juga bisa kita dapatkan dari tabel 2 pada kolom *correctly classified instances* dan *incorrectly classified instances*. *Mean absolute error* juga merupakan kolom yang menyediakan informasi

rata-rata eror yang ada pada beberapa jenis algoritma ketika membangun model klasifikasi untuk 4 algoritma yang tercantum dalam tabel 2.

Tabel 2. Hasil keseluruhan evaluasi dari kinerja beberapa algoritma

Data Set	Algoritma	Mode Tes	*Correctly Classified Instances*		*Incorrectly Classified Instances*		*Mean Absolute Error*
			Angka	%	Angka	%	
E-Coli	Naïve Bayes	*Cross-Validation*	286	85.12 %	50	14.89 %	0.0434
	Lazy- IBK	*Cross-Validation*	270	80.36 %	66	19.65 %	0.0535
	Zero-R	*Cross-Validation*	143	42.56 %	193	57.45 %	0.1829
	Tree- J48	*Cross-Validation*	283	84.23 %	53	15.78 %	0.0486
	Naïve Bayes	*Precentage - Split*	94	82.46 %	20	17.55 %	0.0533
	Lazy- IBK	*Precentage - Split*	94	82.46 %	20	17.55 %	0.0499
	Zero-R	*Precentage - Split*	44	38.6 %	70	61.41 %	0.1858
	Tree- J48	*Precentage - Split*	90	78.95 %	24	21.06 %	0.0621
Yeast	Naïve Bayes	*Cross-Validation*	855	57.62 %	629	42.39 %	0.1046
	Lazy- IBK	*Cross-Validation*	776	52.3 %	708	47.71 %	0.096
	Zero-R	*Cross-Validation*	463	31.2 %	1021	68.81 %	0.1555
	Tree- J48	*Cross-Validation*	747	50.34 %	737	49.67 %	0.1151
	Naïve Bayes	*Precentage - Split*	313	61.99 %	192	38.02 %	0.1036
	Lazy- IBK	*Precentage - Split*	264	52.28 %	241	47.73 %	0.0963
	Zero-R	*Precentage - Split*	160	31.69 %	345	68.32 %	0.1556
	Tree- J48	*Precentage - Split*	268	53.07 %	237	46.94 %	0.1112

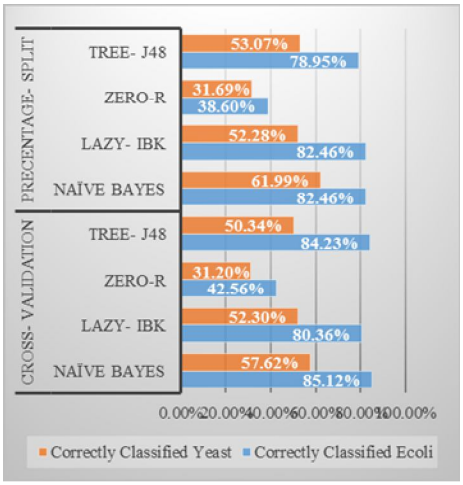
Mode tes *Precentage-split* yang ada pada table 2 dalam makalah ini menggunakan nilai pembagian jumlah data training dan tes sesuai dengan nilai *default* yang disediakan yaitu sebesar 34% untuk data *training* dan 66% untuk data tes. Nilai prosentasi pada kolom persen didaptkan dari hasil

 ISSN: 1411-3201

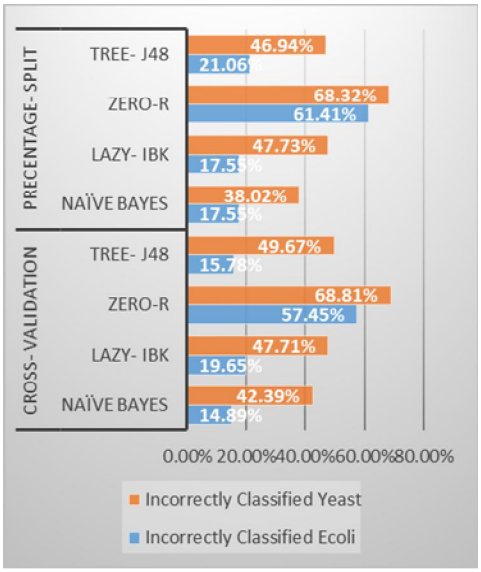
nilai pada kolom angka dibagi dengan total baris data pada data set kmudian dikalikan dengan 100. Hasil evaluasi dicantumkan pada tabel 2.

Jika dilihat secara keseluruhan pada tabel 2, tidak ada nilai akurasi yang mencapai angka 90% Pada kolom *Correctly classified instances* maupun *Incorrectly classified instances.* Angka paling tinggi yang bisa dicapai adalah 85,12% pada algoritma *naïve bayes* untuk data set ecoli yaitu terdiri dari 286 *instances* yang terklasifikasi benar dari 336 data keseluruhan, mencapai nilai *Mean absolute error* sebesar 0,0434. Mode tes yang digunakan untuk akurasi tertinggi tersebut adalah *cross-validation.*

Secara otomatis algoritma *naïve bayesian* yang digunakan untuk mengolah data set ecoli memiliki nilai yang terendah untuk *incorrectly classified instances* sebesar 14,89% dimana klasifikasi data salah hanya sebesasr 50 *instances* dari total keseluruhan data set sebanyak 336 *instances.*



Gambar 1. Hasil perbandingan nilai akurasi klasifikasi data benar



Gambar 2. Hasil perbandingan nilai akurasi klasifikasi data salah

Pada data set Yeast hanya naïve bayes yang mencapai nilai akurasi tertinggi pada mode tes *percentage-split* sebesar 61,99% yaitu sebanyak 313 data terklasifikasi benar dari total 505 *instances.* Sehingga memiliki data terklasifikasi salah yang paling kecil untuk data set yeast yaitu sebesar 38,02%.

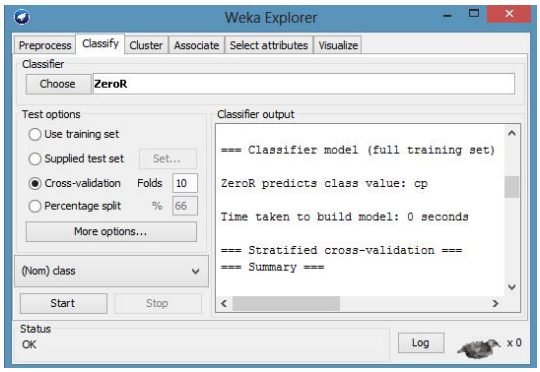
Dari dua grafik yang tertera pada gambar 1 dan gambar 2 memperlihatkan perbandingan nilai akurasi dari beberapa algoritma yang digunakan untuk mengolah data ecoli dan yeast. Pada gambar 1 terlihat bahwa *naïve bayesian* memiliki nilai klasifikasi data benar yang paling tinggi diantara algoritma yang lain untuk data set ecoli maupun data set yeast.

Informasi yang didapat dari grafik pada gambar2 memperlihatkan bahwa algoritma zero-R memiliki nilai klasifikasi data salah paling besar yaitu 68,32% untuk data set yeast dan 61,41% untuk ecoli pada mode test *percentage-split.*

Pada mode tes *Crsoss-validation*, algoritma zero-R juga memiliki nilai akurasi rendah dimana nilai klasifikasi data salah mencapai angka tertinggi yaitu 68,81% pada data set yeast dan 57,45% pada data set ecoli.

Tabel 3. Waktu yang dibutuhkan untuk membangun model

Mode Tes	Algoritma	EColi (Detik)	Yeast (Detik)
Cross-Validation	Naïve Bayes	0.14	0.01
	Lazy- IBK	0.01	0
	Zero-R	0	0
	Tree- J48	0.13	0.09
Precentage- Split	Naïve Bayes	0.01	0.01
	Lazy- IBK	0.01	0
	Zero-R	0	0
	Tree- J48	0.03	0



Gambar 3. Hasil *output* waktu yang digunakan untuk membangun model.

 ISSN: 1411-3201

Dari data tabel 3 dapat diketahui informasi mengenai waktu yang dibutuhkan untuk membangun model pada beberapa algoritma klasifikas. Satuan waktu yang digunakan adalah detik. Mode tes yang digunakan tetap dibagi dua yaitu *Cross-validation dan Precentage-split.*

Gambar 3 adalah salah satu hasil evaluasi untuk algoritma ZeroR yang menggunakan mode tes *Cross-validation* pada data set EColi. Output dari jendela *classifier output* memberikan catatan waktu 0 detik untuk kriteria waktu yang dibutuhkan untuk membangun model.

Algoritma ZeroR memiliki waktu Time to build model yang sangat cepat untuk kedua data set ecoli dan yeast dan mempunyai waktu *Time to build model* di kedua mode tes yaitu sebesar 0 detik.

Kesimpulan dan Saran

Berdasarkan data hasil evaluasi kinerja dari beberapa algoritma klasifikasi yaitu: *Naïve Bayesian, Lazy-IBK, Zero-R,* dan *Decision Tree-J48* dapat disimpulkan bahwa *Naïve Bayesian* memiliki kinerja yang paling baik dalam hal akurasi. Hal tersebut dapat dibuktikan dari nilai *Correctly classified instances* pada data set ecoli mencapai angka prosentase tertinggi sebesar 85,12% pada mode tes *cross-validation.* Begitu juga pada mode tes percentage-split, *naïve Bayesian* mencapai prosentase tertinggi sebesar 82,46%.

Untuk kategori *Time to build model*, algoritma *Zero-R* memilik waktu tercepat 0 detik pada dua data set ecoli dan yeast untuk dua jenis mode tes *cross-validation* maupun *percentage-split.*

Daftar Pustaka

[1] Han, J., & Kamber, M., 2006, *Data Mining: Concepts and Techniques 2e*, Morgan Kaufmann Publishers, San Francisco.
[2] Larose, D.T, 2005, Discovering Knowledge in Data: An Introduction to Data Mining. John Willey & Sons, Inc.
[3] Sheu, Jyh-Jian, May 2008, *An Efficient Two-phase Spam Filtering Methode Based on E-mails categorization.* InternationalJournal of Network Security, Vol. 8, No. 3, PP.334-343,Taiwan.
[4] S. Youn, D. Mcleod, *A,* 2006, *Comparative Study for Email Classification*. Proceedings of International Joint Conferences on Computer, Information, System Sciencesand Engineering, Bridgeport, CT.
[5] Tan, P. N., Steinbach, M., & Kumar, V., 2006, *Introduction to Data Mining*, Pearson Education, Boston.
[6] UCI Machine Learning Repository, 3 Februari 2014, http://archive.ics.uci.edu/ml/

Biodata Penulis

Sulidar Fitri, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika, STMIK AMIKOM Yogyakarta, lulus tahun 2010. Tahun 2012 memperoleh gelar *Master of Science* (M.Sc) dari *Biomedical Informatics Department of Graduate Program Asia University Taiwan*. Saat ini penulis terdaftar sebagai Staf Pengajar di STMIK AMIKOM Yogyakarta. Aktif mengajar sebagai dosen dengan disiplin ilmu yang digeluti adalah Sistem Basis Data, Data Mining, Statistik, dan Struktur Data.