Penerapan Algoritma ID3 dan *Rough Set* untuk Data Tidak Lengkap

Winda Aprianti
Teknik Informatika, Politeknik Negeri Tanah Laut
E-mail: winda.ap17@gmail.com

Abstrak – *Kesalahan prosedur manual entry data, pengukuran, dan peralatan membuat sebagian besar database mengalami permasalahan ketidaklengkapan, seperti missing value. Pada penelitian ini, dataset cuaca tidak lengkap ditangani dengan 3 algoritma klasifikasi, yaitu algoritma klasifikasi 1, 2, dan 3. Algoritma klasifikasi 1 adalah algoritma ID3 dengan penghapusan record yang memiliki missing value, sedangkan algoritma klasifikasi 2 adalah algoritma ID3 dengan penggantian missing value dengan nilai rata-rata dari atribut yang bersesuaian dan memiliki kelas keputusan yang sama, serta algoritma klasifikasi 3 yang merupakan pendekatan rough set yang langsung mempelajari rules dari dataset tidak lengkap. Tujuan penelitian ini adalah mengetahui performansi dari Algoritma 1, 2, dan 3 berdasarkan jumlah rules yang dihasilkan dan akurasi rules. Hasil penelitian menunjukkan bahwa algoritma klasifikasi 1, 2, dan 3 masing-masing menghasilkan 3 rules, 7 rules, serta 10 certain rules dan 16 possible rules, dengan tingkat akurasi 40%, 60%, dan 80%. Hal ini berarti algoritma klasifikasi 3 merupakan algoritma klasifikasi dengan tingkat akurasi tertinggi.*

Kata Kunci *— Akurasi rules, Algoritma ID3, Dataset Tidak Lengkap, Klasifikasi, Missing Value, Rough Set.*

Abstract – *The failure of procedure manual data entry, measurements, and equipment made most of database has incompleteness, such as missing value. In this study, the incomplete meteorological dataset is handled by three classification algorithms, namely 1th classification algorithm, 2nd classification algorithm, and 3rd classification algorithm. 1th classification algorithm is ID3 algorithm with removal of records that have missing value, while 2nd classification algorithm is ID3 algorithm with replacing missing value with average of attribute that have same decision class, and 3rd classification algorithm is rough set approach which directly learn rules of the incomplete dataset. The purpose of this research is to know performance of 1th algorithm, 2nd algorithm, and 3rd algorithm based on number of rules generated and accuracy of the rules. The results showed that 1th classification algorithm, 2nd classification algorithm, and 3rd classification algorithm respectively generate 3 rules, 7 rules, and 10 certain rules and 16 possible rules, with an accuracy rate of 40%, 60%, and 80%. This means that 3rd classification algorithm is classification algorithm with the highest accuracy.*

Keywords *—Accuracy of the rules, ID3 Algorithm, Incomplete Dataset, Classification, Missing Value, Rough Set.*

1. PENDAHULUAN

Perkembangan teknologi *database* yang pesat membuat penyimpanan data dari berbagai sumber dapat dilakukan dengan mudah dan cepat. Akibatnya volume data yang dihasilkan setiap hari meningkat sehingga menganalisis data tersebut menjadi kebutuhan penting. Sebagian besar database di dunia nyata tidak dapat dihindari dari ketidaklengkapan, dalam hal nilai-nilai yang hilang atau salah, disebut dengan *missing value*. Berbagai alasan yang berbeda mengakibatkan ketidaklengkapan dalam data. Contohnya kesalahan prosedur manual *entri* data, pengukuran yang salah, kesalahan peralatan, dan banyak lainnya [1]. Dataset yang tidak lengkap dapat berubah menjadi data yang lengkap dengan berbagai cara, misalnya pada [2] objek dengan nilai yang tidak

diketahui akan dihapus sebelum proses learning dimulai, sedangkan pada [3] *record* yang mengandung *missing values* diganti dengan nilai rata-rata dari atribut lain.

Dataset yang tidak lengkap dalam klasifikasi dapat diproses secara langsung dengan cara tertentu untuk mendapatkan *rules*. Penelitian [4] mengusulkan pendekatan *rough set* untuk langsung mempelajari *rules* dari dataset tidak lengkap tanpa menebak atribut yang tidak diketahui nilai-nilainya.

Cuaca memiliki peran penting dalam kehidupan manusia, seperti dalam bidang kesejahteraan sosial dan ekonomi, pertanian, penanganan bencana, dan keuangan [5]. Jadi, prediksi cuaca perlu dilakukan untuk melakukan perencanaan di berbagai bidang tersebut. Salah satu algoritma yang sering digunakan untuk prediksi cuaca adalah algoritma *Iterative Dichotomicer* 3 (ID3). Hal ini dikarenakan algoritma ID3 merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Selain itu, *rules* yang diperoleh dari pohon keputusan dapat dipahami dengan mudah.

Pada penelitian ini, akan dilakukan penerapan algoritma ID3 dengan penghapusan *missing value*, algoritma ID3 dengan penggantian *missing value* dengan nilai rata-rata, serta pendekatan *rough set* pada dataset cuaca yang mengandung *missing value*. *Rules* yang diperoleh dari ketiga pendekatan tersebut kemudian dibandingkan dengan melihat akurasinya..

2. METODE PENELITIAN

Pada penelitian ini akan digunakan data sekunder cuaca di Perak, Surabaya, Jawa Timur dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). Atribut dari data yang digunakan adalah temperatur, kelembaban, tekanan dan kecepatan angin sebagai atribut pendukung, serta curah hujan sebagai atribut keputusan. *Dataset* tidak lengkap diperoleh dengan cara menghilangkan nilai atribut pendukung dari beberapa objek secara acak, untuk selanjutnya disebut sebagai *incomplete decision table*. *Dataset* dibagi menjadi data pelatihan dan data uji.

Rules dengan menggunakan algoritma ID3 diperoleh dengan cara menghitung *Entropy* dan *Gain* dari setiap atribut, kemudian memilih atribut dengan nilai *Gain* tertinggi. *Entropy* dan *Gain* dihitung menggunakan Persamaan (1) dan (2).

$$Entropy(S) = \sum_{i=1}^{c} -p_i \, log_2 \, p_i \quad \text{..... (1)}$$

$$Gain(S, A) = Entropy(S) - \sum_{v \, \epsilon \, Value \, (A)} \frac{|S_v|}{|S|} \, Entropy \, (S_v) \quad \text{..... (2)}$$

Sedangkan pendekatan *rough set* untuk langsung mempelajari aturan-aturan dari dataset yang tidak lengkap tanpa menebak nilai atribut yang tidak diketahui diusulkan oleh Kryszkiewicz [4]. Hal tersebut dilakukan dengan cara mendefinisikan *similarity relation* antara objek-objek untuk atribut subset *A* sebagai berikut.

$$SIM\ (A) = \{(x,y) \in U \times U \mid a \in A,\ a(x) = a(y) \text{ atau } a(x) = * \text{ atau } a(y) = *\} \quad \text{..... (3)}$$

Rough set merupakan pendekatan konsep *vague* oleh pasangan konsep yang tepat, disebut *lower* dan *upper approximation*. Misal *X* adalah subset dari semesta *U* dan *B* adalah sebarang *subset* dari atribut *A*. *Lower* dan *upper approximation B* terhadap *X,* masing-masing didefinisikan pada Persamaan (4) dan Persamaan (5).

$$\underline{B}X = \{x \mid x \in U,\ B(x) \subseteq X\} \quad \text{.....(4)}$$

dan

$$\overline{B}X = \{x \mid x \in U \text{ dan } B(x) \cap X \neq \emptyset\} \quad \text{.....(5)}$$

Rules diperoleh dengan cara menerapkan *rough set* pada *incomplete decision table*. Pertama kita membentuk subset dari setiap atribut, kemudian mencari *lower* dan *upper approximation*. Kita

akan memperoleh *certain rules* dari *lower approximation*, sedangkan *possible rules* diperoleh dari *upper approximation*.

Untuk mengetahui tingkat akurasi dari masing-masing *rules* yang diperoleh dari algoritma di atas, maka dilakukan perbandingan kelas hasil prediksi terhadap data uji dengan kelas klasifikasi yang sebenarnya. Selanjutnya, dilakukan perbandingan tingkat akurasi dari ketiga.

3. HASIL DAN PEMBAHASAN

3.1 Algoritma Klasifikasi 1

Algoritma klasifikasi pertama yang digunakan adalah algoritma ID3 dengan penghapusan *missing value*. Pada algoritma ini, *record* yang memiliki *missing value* di satu atau lebih atribut akan dihilangkan, sehingga diperoleh *dataset* baru berupa *dataset* cuaca yang lengkap. Selanjutnya algoritma ID3 diterapkan pada dataset baru yang terbentuk untuk menghasilkan *rules*.

3.2 Algoritma Klasifikasi 2

Algoritma klasifikasi kedua yang digunakan adalah algoritma ID3 dengan penggantian missing value dengan nilai rata-rata. Sebelum menerapkan algoritma ID3, dataset cuaca tidak lengkap akan diubah menjadi dataset lengkap dengan cara mengganti *missing value* dengan nilai rata-rata pada atribut yang bersesuaian dan memiliki kelas keputusan yang sama.

3.3 Algoritma Klasifikasi 3

Ide dasar algoritma klasifikasi ketiga mengacu pada kajian [4] tentang pendekatan *rough set* pada sistem informasi tidak lengkap dan kajian [6] tentang penggunaan *fuzzy rough set* terhadap data kuantitatif tidak lengkap. Berdasarkan kedua kajian tersebut, dapat dibentuk algoritma klasifikasi sebagai berikut.

Langkah 1: Partisi himpunan objek-objek ke dalam *subset disjoint* menurut label kelas.
Langkah 2: Transformasi nilai kuantitatif menjadi kategorikal. Jika *Objek(i)* memiliki nilai yang hilang untuk atribut A_j, tulis tetap dengan nilai hilang (*).
Langkah 3: Temukan *equivalence class* dari atribut tunggal.
Langkah 4: Inisialisasi $q = 1$, dimana q digunakan untuk menghitung jumlah atribut saat diproses untuk *lower* dan *upper approximations.*
Langkah 5: Hitung *lower* dan *upper approximations* dari setiap subset B dengan q atribut untuk setiap kelas X_l.
Langkah 6: Tetapkan $q = q+1$ dan ulangi langkah 5-7 sampai $q > m$.
Langkah 7: Derivasi *certain rules* dari *lower approximation* dan *possible rules* dari *upper approximation* pada setiap subset B.
Langkah 8: Hapus *certain* dan *possible rules* dengan kondisi bagian yang lebih spesifik daripada *certain* dan *possible rules* lainnya.
Langkah 9: *Output certain* dan *possible rules*.

3.4 Pengujian Algoritma dan Analisis Hasil

Untuk menguji algoritma di atas, digunakan dataset pada [7]. Dataset tersebut ditampilkan pada Tabel 1.

Tabel 1. Dataset Cuaca Tidak Lengkap

Obj	A	B	C	D	CH
1.	27.7	92.1	1006.9	*	1
2.	26.3	94.1	*	4.8	70.1
3.	*	93.9	1006.4	7.1	59.9
4.	27.5	94.3	1007.3	3.2	37.1
5.	24	96.9	1008	5.5	8.9
6.	26.1	*	1009.1	1.6	17
7.	*	95.1	1012.7	6.4	33
8.	27.9	67.2	1011.7	10.3	0
9.	29	74.5	1012	12.6	0
10.	31.9	82.4	*	6.8	0
11.	29.2	88.2	1008.1	6.6	0
12.	29.1	*	1008.5	7.2	38.1
13.	28.4	87.6	1007.3	9.7	0
14.	25.2	97.1	1007.4	5	37.1
15.	*	95.8	1008.8	4	4.1

Keterangan:
Obj = Objek
A = Temperatur
B = Kelembaban
C = Tekanan Udara
D = Kecepatan Angin
CH = Curah Hujan

Sebelum menerapkan algoritma *rough* set, semua atribut pada Tabel 1 diubah ke dalam bentuk kategorikal dengan *missing value* disimbolkan *. Untuk atribut temperatur, ketika temperatur kurang dari 26.5 diganti menjadi sejuk, temperatur yang berada di antara 26.5 dan 29 diganti menjadi normal, dan temperatur yang lebih dari 29 diganti menjadi panas. Untuk atribut kelembaban, ketika kelembaban kurang dari 68 diganti menjadi kering, kelembaban yang berada di antara 68 dan 78 diganti menjadi lembab, dan kelembaban yang lebih dari 78 diganti menjadi basah. Untuk atribut tekanan, jika tekanan kurang dari 1008 diganti menjadi rendah, jika tekanan berada di antara 1008 dan 1013 diganti menjadi normal, dan jika tekanan yang lebih dari 1013 diganti menjadi tinggi. Untuk atribut kecepatan angin, jika kecepatan kurang dari 4 diganti menjadi lambat, jika kecepatan berada di antara 4 dan 8 diganti menjadi normal, dan kecepatan yang lebih dari 8 diganti menjadi kencang. Sedangkan untuk atribut keputusan atau curah hujan (CH) diberikan kategori Ringan (R) jika CH kurang dari 20 dan kategori Lebat (L) jika CH lebih dari 20.

Dataset yang terdiri dari 15 objek dibagi menjadi 2, yaitu data pelatihan dan data uji. Objek 1 sampai objek 10 digunakan sebagai data pelatihan, sedangkan objek 11 sampai objek 15 digunakan sebagai data uji.

Pada pengujian menggunakan algoritma klasifikasi 1, dataset baru yang terbentuk terdiri dari 4 *record* dan menghasilkan 3 buah *rules* seperti yang ditunjukkan pada Tabel 2.

Tabel 2. *Rules* Hasil Algoritma Klasifikasi 1

No.	***Rules***
1.	Jika (D = lambat) maka CH = L
2.	Jika (D = normal) maka CH = R
3.	Jika (D = kencang) maka CH = R

Pada pengujian menggunakan algoritma klasifikasi 2, dataset baru tetap berjumlah 10 *record*, tetapi telah menjadi dataset lengkap dikarenakan adanya proses penggantian *missing value* dengan nilai rata-rata dari atribut yang bersesuaian dan memiliki kelas keputusan yang sama. Pengujian algoritma klasifikasi 2 pada dataset baru yang terbentuk menghasilkan 7 *rules* seperti yang ditunjukkan pada Tabel 3.

Tabel 3. *Rules* Hasil Algoritma Klasifikasi 2

No.	***Rules***
1.	Jika (A = panas) maka CH = R
2.	Jika (A = sejuk) dan (C = rendah) maka CH = R
3.	Jika (A = sejuk), (C = normal), dan (D = normal) maka CH = L
4.	Jika (A = sejuk), (C = normal), dan (D = lambat) maka CH = R
5.	Jika (A = normal) dan (D = lambat) maka CH = L
6.	Jika (A = normal), (C = normal), dan D = normal) maka CH = L
7.	Jika (A = normal) dan (D = kencang) maka CH = R

Pengujian menggunakan algoritma klasifikasi 3 yang langsung diterapkan pada dataset cuaca tidak lengkap menghasilkan 10 *certain rules* yang diperoleh dari *lower approximation* dan 16 *possible rules* yang diperoleh dari *upper approximation* seperti yang ditunjukkan pada Tabel 4.

Tabel 4. *Rules* Hasil Algoritma Klasifikasi 3

No.	***Rules***
	Certain Rules
1.	Jika (B = kering) maka CH = R
2.	Jika (B = lembab) maka CH = R
3.	Jika (D = kencang) maka CH = R
4.	Jika (A = sejuk) dan (D = lambat) maka CH = R
5.	Jika (C = normal) dan (D = lambat) maka CH = R
6.	Jika (A = panas), (B = basah), dan (C = rendah) maka CH = R
7.	Jika (A = normal), (B = basah), dan (C = normal) maka CH = L
8.	Jika (A = panas), (B = basah), dan (D = normal) maka CH = R
9.	Jika (A = normal), (C = normal), dan (D = normal) maka CH = L
10.	Jika (A = sejuk), (C = normal), dan (D = normal) maka CH = L
	Possible Rules
11.	Jika (A = sejuk) maka CH = R
12.	Jika (A = sejuk) maka CH = L
13.	Jika (A = normal) maka CH = R
14.	Jika (A = normal) maka CH = L

15. Jika (A = panas) maka CH = R
16. Jika (A = panas) maka CH = L
17. Jika (B = basah) maka CH = R
18. Jika (B = basah) maka CH = L
19. Jika (C = rendah) maka CH = R
20. Jika (C = rendah) maka CH = L
21. Jika (C = normal) maka CH = R
22. Jika (C = normal) maka CH = L
23. Jika (D = lambat) maka CH = R
24. Jika (D = lambat) maka CH = L
25. Jika (D = normal) maka CH = R
26. Jika (D = normal) maka CH = L

Hasil *rules* pada Tabel 2, 3, dan 4 digunakan untuk memprediksi curah hujan pada data uji. Hasil prediksi ini ditunjukkan pada Tabel 5.

Tabel 5. Hasil Prediksi Data Uji

Objek	Algoritma Klasifikasi 1	Algoritma Klasifikasi 2	Algoritma Klasifikasi 3
11.	R	R	R
12.	R	R	R
13.	R	R	R
14.	R	R	L
15.	L	-	R

Algoritma klasifikasi 3 memprediksi klasifikasi dari objek 14 dengan cara menghitung *possible rules* untuk atribut A = dingin, B = basah, C = rendah, dan D = normal yang memiliki 2 keputusan yang berbeda, yaitu CH = R dan CH = L. Dalam kasus ini kita memilih CH = L karena pada data pelatihan objek yang memiliki keputusan CH = L lebih banyak dibandingkan objek yang memiliki kelas keputusan CH = R. Namun, untuk mendapatkan hasil prediksi yang lebih efektif, aturan untuk memilih *possible rules* perlu didefinisikan lebih lanjut pada kajian berikutnya.

Hasil perbandingan Tabel 5 dengan objek 11 sampai objek 15 pada Tabel 1 menunjukkan bahwa akurasi dari algoritma klasifikasi 1, 2, dan 3 masing-masing adalah 40%, 60%, dan 80%. Penghapusan *record* yang memiliki *missing value* membuat *rules* yang dihasilkan oleh algoritma klasifikasi 1 kurang merepresentasikan data. Selain itu, *rules* yang dihasilkan oleh algoritma klasifikasi 1 hanya bisa digunakan apabila atribut kecepatan angin (D) diketahui nilainya. Sedangkan untuk *rules* yang dihasilkan oleh algoritma klasifikasi 2 belum bisa menangani semua kondisi. Hal ini ditunjukkan oleh ketidakmampuan *rules* hasil algoritma klasifikasi 2 dalam memprediksi objek 15. Ketidakmampuan ini dapat diatasi oleh *rules* yang dihasilkan oleh algoritma klasifikasi 3. Performansi algoritma klasifikasi 3 juga lebih baik dalam hal akurasi yang memperoleh akurasi paling tinggi dibandingkan algoritma klasifikasi 1 dan 2.

4. SIMPULAN

Pada penelitian ini telah dibahas mengenai 3 buah algoritma klasifikasi untuk mendapatkan *rules* dari dataset cuaca tidak lengkap. Algoritma klasifikasi 1 melakukan penghapusan *record* yang memiliki *missing value* sebelum menerapkan algoritma ID3, yaitu membentuk akar dan cabang berdasarkan perhitungan nilai *gain* tertinggi. Algoritma klasifikasi 2 juga menerapkan algoritma ID3, tetapi *missing value* diganti dengan nilai rata-rata dari atribut yang bersesuaian dan memiliki

kelas keputusan yang sama. Sedangkan algoritma klasifikasi 3 yang mendapatkan *rules* secara langsung menggunakan pendekatan *rough set*. *Rules* diperoleh dengan cara membentuk subset dari setiap atribut, kemudian mencari *lower* dan *upper approximation*. Selanjutnya akan diperoleh *certain rules* dari *lower approximation*, sedangkan *possible rules* diperoleh dari *upper approximation*. Langkah terakhir adalah menghapus *certain* dan *possible rules* yang lebih spesifik dari *rules* lainnya.

Penerapan algoritma klasifikasi 1, 2, dan 3 pada dataset cuaca tidak lengkap dengan 5 atribut menghasilkan 3 *rules*, 7 *rules*, serta 10 *certain rules* dan 16 *possible rules*. Penggunaan masing-masing *rules* untuk klasifikasi data uji menunjukkan tingkat akurasi 40%, 60%, dan 80%. Hal ini berarti *rules* hasil algoritma klasifikasi dengan pendekatan *rough set* pada dataset tidak lengkap lebih efektif untuk memprediksi cuaca di waktu mendatang.

5. SARAN

Penelitian selanjutnya akan ditingkatkan tentang cara menetapkan perhitungan *rules* hasil algoritma klasifikasi *rough set* terhadap klasifikasi data uji, terutama *possible rules*.

DAFTAR PUSTAKA

[1] Sadiq, A.T, Dualmi, M.G., dan Shaker, A.S. 2013. Data Missing Solution Using Rough Set Theory and Swarm Intelligence. *International Journal of Advanced Computer Science and Information Technology (IJACSIT)* Vol. 2, No. 3, 2013, Page: 1-16, ISSN: 2296-1739.

[2] Chmielewski, M.R., Gryzmala-Busse, J.W., Peterson, N.W., dan Than, Soe. 1993. The rule induction system LERS – A version for personal computers. *Foundations of Computing and Decision Sciences*, 18, 181–212.

[3] Iqbal, M., Mukhlash, I., dan Astuti, H.M. 2013. The Comparison of CBA Algorithm and CBS Algorithm for Meteorological Data Classification. *Information Systems International Conference (ISICO),* 2-4 December 2013.

[4] Kryszkiewicz, M. 1998. Rough Set Approach to Incomplete Information Systems. *Information Science,* Vol . 112, No. 1, pp. 39-49.

[5] National Council of Applied Economic Research. 2010. *Impact Assessment and Economic Benefits of Weather and Marine Services*. Diakses dari http://www.ncacr.org pada tanggal 11 April 2016.

[6] Hong, T., Tseng, L., dan Cien, B. 2009. Mining from Incomplete Quantitative by Fuzzy Rough Sets. *Expert Systems with Application* DOI:10.1016/j.eswa.2009.08.002.

[7] Aprianti, W. dan Mukhlash, I. 2014. The Application of Rough Set and Fuzzy Rough Set Based Algorithm to Classify Incomplete Meteorological Data. *2014 International Conference on Data and Software Engineering*, pp. 177-182, ISBN: 978-1-4799-8177.