Average Weight Information Gain Untuk Menangani Data Berdimensi Tinggi Menggunakan Algoritma C4.5

Joko Suntoro, Cahya Nurani Indah
Program Studi Magister Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Jl. Imam Bonjol 205-207, Lt 2, Kota Semarang
E-mail: joko@jokosuntoro.com, cahya.nurani28@gmail.com

Masuk: 15 Februari 2017; Diterima: 15 Februari 2017

Abstract. *In the recent decades, a large data are stored by companies and organizations. In terms of use, big data will be useless if not processed into information according to the usability. The method used to process data into information is called data mining. The problem in data mining especially classification is data with a number of attributes that many and each attribute are irrelevant. This study proposes attribute weighting method using weight information gain method, then the attribute weights calculates the average value. Having calculated the average value of the attribute selection, the selected attributes are those with a value weights above average value. Attributes are selected then performed using an algorithm C4.5 classification, this method is named Average Weight Information Gain C4.5 (AWEIG-C4.5). The results show that AWEIG-C4.5 method is better than C4.5 method with the accuracy of the average value of each is 0.906 and 0.898.*
Keywords: *data mining, high dimensional data, weight information gain, C4.5 algorithm*

Abstrak. *Dalam beberapa dekade terakhir, data yang besar disimpan oleh perusahaan dan organisasi. Dari segi penggunaan, data besar tersebut akan menjadi tidak berguna jika tidak diolah menjadi informasi yang sesuai dengan kegunaan. Metode yang digunakan untuk mengolah data menjadi informasi adalah data mining. Masalah dalam data mining khususnya klasifikasi adalah data dengan jumlah atribut yang banyak atau dalam bahasa komputer disebut data berdimensi tinggi. Pada penelitian ini diusulkan metode pembobotan atribut menggunakan metode weight information gain, kemudian bobot atribut tersebut dihitung nilai rata-rata. Setelah dihitung nilai rata-rata dilakukan pemilihan atribut, atribut yang dipilih adalah atribut dengan nilai bobot di atas nilai rata-rata. Atribut yang terpilih kemudian dilakukan klasifikasi menggunakan algoritma C4.5, metode ini diberi nama Average Weight Information Gain C4.5 (AWEIG-C4.5). Hasil penelitian menunjukkan metode AWEIG-C4.5 lebih baik daripada metode C4.5 dengan nilai rata-rata akurasi masing-masing adalah 0,906 dan 0,898. Dari uji paired t-Test terdapat perbedaan signifikan antara metode AWEIG C4.5 dengan metode C4.5.*
Kata Kunci: *data mining, data berdimensi tinggi, weight information gain, algoritma C4.5*

1. Pendahuluan

1.1. Latar Belakang Penelitian

Dalam beberapa dekade terakhir, data yang besar disimpan oleh perusahaan dan organisasi. Data tersebut berasal dari beberapa format, mulai dari *text*, gambar, suara, email, pembacaan sensor, dan sebagainya. Dari segi penggunaan, data besar tersebut akan menjadi tidak berguna jika tidak diolah menjadi informasi yang sesuai dengan kegunaan. Metode yang digunakan untuk mengolah data menjadi informasi adalah *data mining* (Jiawei Han, Micheline Kamber, 2012).

Data mining sudah diterapkan dalam banyak bidang antara lain: estimasi proyek perangkat lunak untuk perusahaan manufaktur (Chou & Wu, 2013), prediksi terjadinya gempa bumi (Moustra, Avraamides, & Christodoulou, 2011), klasifikasi *web spam* (Fdez-Glez et al., 2015), deteksi klastering sel kanker pada bagian payudara (Oliver et al., 2012) dan penyusunan tata letak barang pada supermarket melalui *asosociation rule* berdasarkan kebiasaan pembelian konsumen (Cil, 2012). Secara umum kegunaan dari data mining adalah estimasi, prediksi, klasifikasi, klastering dan asosiasi.

Masalah dalam data mining khususnya klasifikasi adalah data berdimensi tinggi (Jin, Jin, & Qin, 2012). Data berdimensi tinggi menyebabkan ukuran *dataset* menjadi lebih besar, jumlah atribut yang banyak maupun jumlah data sampel yang besar (Bennasar, Hicks, & Setchi, 2015). Data dengan banyak atribut menyebabkan performa algoritma klasifikasi menjadi rendah (Sáez, Derrac, Luengo, & Herrera, 2014)(Qian & Shu, 2015).

Salah satu metode yang digunakan untuk mengatasi data berdimensi tinggi adalah seleksi fitur (Wahono, Suryana, & Ahmad, 2014). Seleksi fitur berguna untuk mengurangi ukuran atribut yang besar pada *dataset* (Sebastiani, 2002). Metode seleksi fitur dibagi menjadi tiga teknik yaitu *filter*, *wrapper* dan *hybrid* (G. Chen & Chen, 2015). Teknik *filter* menggunakan relevansi antar atribut berdasarkan sifat intrinsik dari data (Sun et al., 2013), teknik *wrapper* memilih atribut berdasarkan dari evaluasi kinerja *classifier* (Hajek & Michalak, 2013) sedangkan teknik *hybrid* menggabungkan antara teknik *filter* dan teknik *wrapper*.

Metode *chi-square* adalah salah satu metode dari teknik *filter*. Metode *chi-square* pernah digunakan oleh Daliri (Daliri, 2013) untuk melakukan seleksi fitur pada data penyakit parkinson. Metode *chi-square* untuk seleksi fitur berdasarkan pada penghitungan χ^2 statistik, fitur yang dipilih adalah fitur dengan nilai chi-square tertinggi (Verónica Bolón-Canedo, Noelia Sánchez-Maroño, 2015). Pada teknik *filter*, selain metode *chi-square*, penelitian Koprinska et al (Koprinska, Rana, & Agelidis, 2015) menerapkan metode *correlation based feature selection* (CBFS) untuk memprediksi beban listrik. Tujuan dari CBFS adalah atribut yang baik harus saling berkolerasi antara satu atribut dengan atribut lainnya.

Metode *wrapper* yang dikembangkan dengan *local maximization* (LM) dan *floating maximization* (FM) pada algoritma SVM pernah diusulkan oleh Korfiatis et al (Korfiatis, Asvestas, Delibasis, & Matsopoulos, 2013) untuk seleksi fitur diagnosis penyakit tulang sumsum. Subset yang terbaik dari ukuran dipilih dengan LM. Solusi terbaik dengan mengubah ukuran awal subset digunakan FM dengan menggunakan teknik ukuran *floating*. Pada level klasifikasi algoritma SVM digunakan untuk melakukan diskriminasi dari data.

Teknik *filter* lebih cepat digunakan dibandingkan dengan teknik *wrapper* dan teknik *hybrid*, selain itu teknik *filter* meningkat lebih baik dan mudah diterapkan daripada teknik *wrapper* dan teknik *hybrid* (Bolón-Canedo, Porto-Díaz, Sánchez-Maroño, & Alonso-Betanzos, 2014). Pada penelitian ini, digunakan metode *weight information gain*. Metode *weight information gain* masuk ke dalam teknik *filter*. Masing-masing atribut dihitung nilai bobotnya, kemudian dihitung nilai rata-ratanya. Atribut yang dipilih adalah atribut dengan nilai bobot di atas nilai rata-rata.

Model klasifikasi yang digunakan pada penelitian ini adalah algoritma C4.5. Algoritma C4.5 dapat menggambarkan secara eksplisit struktur model, struktur model pada algoritma C4.5 berupa pohon akar, sehingga banyak digunakan dibandingan dengan algoritma lainnya (Thammasiri, Delen, Meesad, & Kasap, 2014). Penelitian Chen et al (K. H. Chen, Wang, Wang, & Angelia, 2014) menunjukkan bahwa hasil evaluasi algoritma C4.5 lebih baik dibandingkan dengan dengan algoritma SVM, SOM dan BPNN untuk klasifikasi pada penyakit kanker. Selain itu menurut Wu et al (Wu et al., 2008) algoritma C4.5 masuk ke dalam 10 besar algoritma terbaik, khususnya algoritma klasifikasi, sehingga dipilih pada penelitian ini.

1.2. Identifikasi Masalah

Dari latar belakang di atas, maka identifikasi masalah pada penelitian ini dapat diidentifikasikan bahwa data dengan banyak atribut dan masing-masing atribut tidak relevan menyebabkan performa algoritma klasifikasi menjadi rendah.

1.3. Rumusan Masalah

Berdasarkan latar belakang dan identifikasi masalah, maka rumusan masalah pada penelitian ini adalah bagaimana peningkatan hasil evaluasi pada penerapan metode *average weight information gain* untuk menangani data dengan banyak atribut.

1.4. Tujuan Penelitian

Tujuan penelitian ini adalah mengembangkan sebuah metode pembobotan atribut menggunakan penghitungan *weight information gain* untuk menangani data dengan banyak atribut, kemudian dihitung nilai rata-rata dari masing-masing bobot atribut, atribut yang dipilih adalah atribut dengan nilai bobot di atas nilai rata-rata, setelah itu diklasifikasi dengan algoritma C4.5 sehingga dapat meningkatkan hasil evaluasi.

2. Tinjauan Pustaka

2.1. *Weight Information Gain*

Weight Information gain (WIG) adalah metode pembobotan tiap variabel yang paling umum dari atribut evaluasi (Verónica Bolón-Canedo, Noelia Sánchez-Maroño, 2015). Untuk menghitung *information gain,* terlebih dahulu harus memahami suatu aturan lain yang disebut *entropy*. Di dalam bidang *Information Theory*, kita sering menggunakan *entropy* sebagai suatu parameter untuk mengukur heterogenitas (keberagaman) dari suatu kumpulan sampel data. Jika kumpulan sampel data semakin heterogen, maka nilai *entropy* semakin besar. Secara matematis, *entropy* dirumuskan dengan persamaan (1), dimana *c* adalah jumlah nilai yang ada pada atribut target (jumlah kelas). Sedangkan p_i menyatakan jumlah sampel untuk kelas *i*.

Setelah mendapatkan nilai *entropy* untuk suatu kumpulan sampel data, maka kita dapat mengukur efektifitas suatu atribut. Ukuran efektifitas ini disebut sebagai *information gain*. Secara matematis, *information gain* dari suatu atribut A, dituliskan pada persamaan (2), dimana nilai *A* adalah atribut, *V* menyatakan suatu nilai yang mungkin untuk atribut *A*, *Value(A)* adalah himpunan nilai-nilai yang mungkin untuk atribut *A*. $|S_v|$ adalah jumlah sampel untuk nilai *v*, |*S*| jumlah seluruh data dan *Entropy(Sv)* adalah *entropy* untuk sampel-sampel yang memiliki nilai *v*.

$$entropy = \sum_{i}^{c} - p_i \log_2 p_i \quad (1)$$

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

2.2. Algoritma C4.5

Sistem yang menghasilkan pengklasifikasian merupakan *tool* paling populer di data mining, salah satunya adalah C4.5 yang merupakan turunan dari CLS (*Concept Learning System*) dan ID3 (*Iterative Dichotomiser 3*). Seperti CLS dan ID3, C4.5 juga menghasilkan pengklasifikasian yang digambarkan melalui *decision tree*, tetapi C4.5 juga menghasilkan pengklasifikasian dalam bentuk rule. Pembentukan *decision tree* dihasilkan dari proses induksi.

Induksi di C4.5 terdiri dari dua fase (Gorunescu, 2011) yaitu: (1) membentuk *initial tree*, menggunakan *training dataset* sampai masing-masing daun (*leaf*) menjadi homogen atau mendekati murni (*pure*), (2) *pruning*, yaitu memangkas cabang (*branches*) yang tidak diperlukan dari *decision tree*. Algoritma C4.5 menggunakan konsep *information gain* atau *entropy reduction* untuk menentukan *split* yang optimal (Larose, 2005). *Split* yang terpilih adalah yang mempunyai nilai *information gain* dan *gain* yang terbesar.

Tahapan dalam membentuk *decision tree* pada algoritma C4.5 adalah (1) menghitung nilai *entropy*, (2) menghitung nilai *Gain Ratio* untuk masing-masing atribut, (3) atribut yang memiliki *Gain Ratio* tertinggi dipilih menjadi akar (*root*) dan atribut yang memiliki nilai *Gain Ratio* lebih rendah dari akar (*root*) dipilih menjadi cabang (*branches*), (4) menghitung lagi nilai *Gain Ratio* tiap-tiap atribut dengan tidak mengikutsertakan atribut yang terpilih menjadi akar (*root*) di tahap sebelumnya, (5) atribut yang memiliki Gain Ratio tertinggi dipilih menjadi cabang (*branches*), (6) mengulangi langkah ke-4 dan ke-5 sampai dengan dihasilkan nilai *Gain* = 0 untuk semua atribut yang tersisa.

2.3. Metodologi Penelitian

Menurut jenis metode, pada penelitian ini digunakan penelitian eksperimen. Menurut Dawson (Dawson, 2011), penelitian eksprimen adalah uji coba yang dikontrol oleh peneliti sendiri untuk melakukan investigasi hubungan kausal (hubungan sebab-akibat). Langkah-langkah penelitian dapat dilihat pada Gambar 1, yang meliputi: (1) Analisa permasalahan dan tinjauan pustaka. Penelitian ini diawali dengan mengumpulkan *survey paper* dan *technical paper* yang berhubungan dengan metode yang diusulkan kemudian mengetahui *state-of-the-art methods* penelitian yaitu data berdimensi tinggi. (2) Pengumpulan *dataset* dan pengolahan data. *Dataset* yang digunakan pada penelitian ini adalah menggunakan *dataset public*. *Dataset public* yang digunakan adalah UCI *dataset repository* antara lain: *bank marketing*, *churn data*, *credit approval*, *ionosphere*, dan *chronic kidney* yang bisa diunduh melalui https://archive.ics.uci.edu/ml/datasets.html. (3) Metode yang diusulkan. Pada tahapan ini dijelaskan metode yang diusulkan pada penelitian. Data berdimensi tinggi dengan jumlah atribut yang banyak akan dilakukan pembobotan atribut dengan metode *weight information gain*, kemudian dihitung nilai rata-ratanya. Atribut yang dipilih adalah atribut dengan nilai bobot di atas nilai rata-rata. *Dataset* dengan atribut baru tersebut kemudian diklasifikasi menggunakan algoritma C4.5. Metode ini dinamakan dengan *average weight information gain* – C4.5 (AWEIG-C4.5). (4) Eksperimen dan pengujian metode. Pada bagian ini metode yang diusulkan (AWEIG-C4.5) akan dibandingkan dengan metode C4.5 biasa. Kemudian dicatat hasil akurasi masing-masing metode dan dipilih hasil yang terbaik. (5) Evaluasi hasil. Setelah dilakukan eksperimen terhadap semua *dataset* dengan metode AWEIG-C4.5 dan metode C4.5 biasa, kemudian hasil tersebut dievaluasi. Dari hasil evaluasi dapat ditarik kesimpulan dari eksperimen.

3. Metode Penelitian

3.1. Pengumpulan *Dataset* dan Pengolahan Data

Dataset yang digunakan pada penelitian ini adalah menggunakan *dataset public*. *Dataset* yang digunakan pada penelitian ini pernah digunakan oleh peneliti sebelumnya dengan topik data berdimensi tinggi. *Dataset public* yang digunakan adalah UCI *dataset repository* antara lain: *bank marketing* (Ispandi & Wahono, 2015), *churn data* (Farquad, Ravi, & Raju, 2014), *credit approval* (Wang & Huang, 2009), *ionosphere* (Zhao, Fu, Ji, Tang, & Zhou, 2011), dan *chronic kidney* (Korfiatis et al., 2013) yang bisa diunduh melalui https://archive.ics.uci.edu/ml/datasets.html.

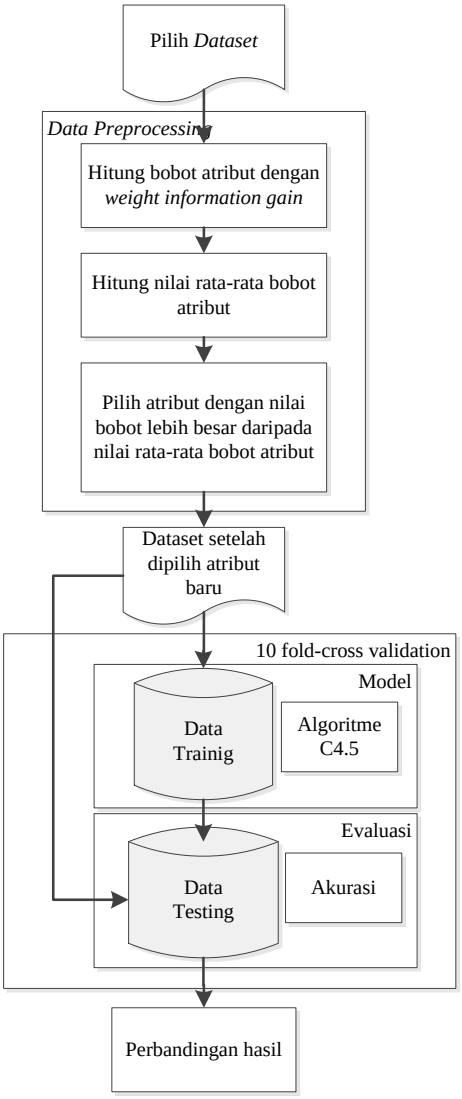
Tabel 1 menunjukkan *dataset* yang digunakan pada penelitian ini. Pada penelitian ini terdapat *dataset* yang mengandung *missing value*. *Dataset* yang mengandung *missing value* diganti dengan nilai rata-rata dari masing-masing atribut (Jiawei Han, Micheline Kamber, 2012).

Tabel 1. *Dataset* yang Digunakan pada Penelitian

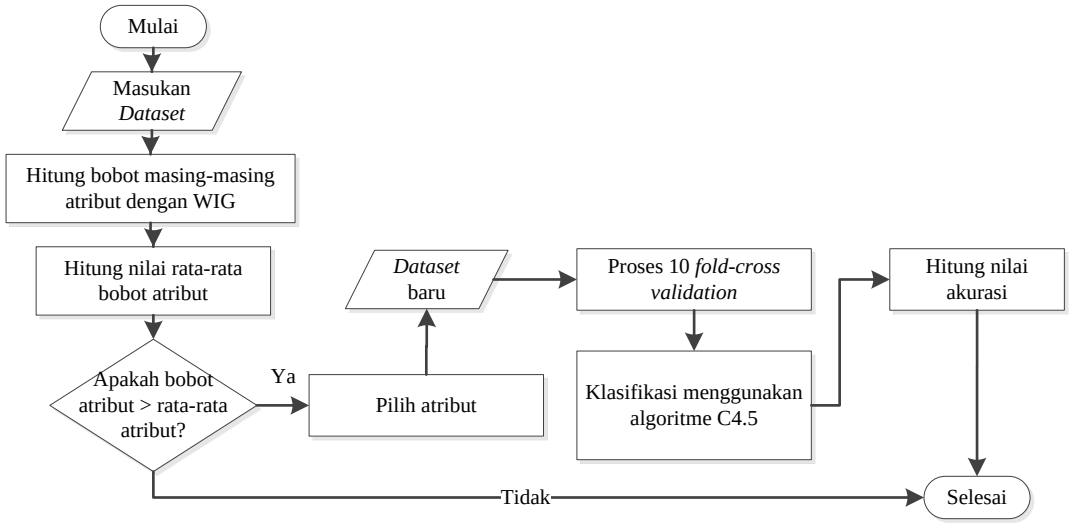
No	*Dataset*	*Jumlah Atribut*	*Jumlah Data*
1	Bank Marketing	17	4.521
2	Chronic Kidney	25	400
3	Churn Data	21	4.975
4	Credit Approval	16	690
5	Ionosphere	35	350

3.2. Metode yang Diusulkan (*Average Weight Information Gain* – C4.5)

Gambar 1 merupakan metode yang diusulkan pada penelitian ini. *Dataset* dengan banyak atribut digunakan pada penelitian ini. *Dataset* dengan banyak atribut tersebut dipecahkan dengan metode fitur seleksi yaitu *weight information gain*. Metode yang diusulkan diberi nama *Average Weight Information Gain* – C4.5 (AWEIG-C4.5). *Flowchart* metode AWEIG-C4.5 dapat dilihat pada Gambar 2.



Gambar 1. Metode yang Diusulkan (AWEIG-C4.5)



Gambar 2. *Flowchart* Metode AWEIG-C4.5

Berikut adalah langkah-langkah metode AWEIG-C4.5 pada penelitian ini: (1) Pilih *dataset*. (2) Hitung bobot masing-masing atribut dengan persamaan (3). (3) Hitung nilai rata-rata bobot atribut (*AVG*) dengan persamaan (4). (4) Pilih atribut dengan nilai bobot lebih besar daripada nilai rata-rata bobot atribut. (5) *Dataset* yang telah dipilih atribut baru tersebut, dibagi dua menjadi data *training* dan data *testing* menggunakan 10 *fold-cross validation*. (6) Klasifikasi menggunakan algoritma C4.5. (7) Hitung nilai akurasi berdasarkan tabel *confusion matrix*.

$$WIG = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (3)$$

$$AVG = \frac{A_1 + A_2 + A_3 + ... + A_N}{N} \quad (4)$$

4. Eksperimen dan Pengujian Metode

Eksperimen ini dilakukan menggunakan program bantu Rapidminer versi 7.2.001, Microsoft Excel 2013 dan XLSTAT 2016. Tabel 2 menunjukkan spesifikasi komputer yang digunakan pada penelitian ini.

Uji *paired t-Test* digunakan untuk mengetahui apakah ada perbedaan signifikan diantara dua model. Dalam uji *paired t-Test* ditetapkan nilai α sebesar 0,05, jika nilai *p-value* lebih besar dari nilai α maka tidak ada perbedaan signifikan antar model. Sedangkan jika nilai *p-value* lebih kecil dari nilai α maka terdapat perbedaan signifikan antar model.

Hasil rekap pengukuran akurasi antara metode C.45 dan AWEIG-C4.5 dapat dilihat pada Tabel 3. Hasil pengukuran akurasi dicatat berdasarkan hasil eksperimen pada *dataset bank marketing*, *churn data*, *credit approval*, *ionosphere*, dan *chronic kidney*. Metode AWEIG-C4.5 lebih baik daripada metode C4.5 dengan nilai rata-rata akurasi masing-masing adalah 0,906 dan 0,898. Tabel 3 menunjukkan hasil pengukuran akurasi pada metode C4.5 dan AWEIG-C4.5 dan Gambar 3 menunjukkan diagram perbandingan akurasi pada metode C4.5 dan AWEIG-C4.5.

Uji *paired t-Test* digunakan agar diketahui apakah ada perbedaan signifikan antara metode C4.5 dan AWEIG-C4.5. Tabel 4 adalah hasil uji *paired t-Test* antara metode C4.5 dan AWEIG-C4.5. Dari hasil uji tersebut didapatkan nilai *p-value* adalah 0,04, ini menunjukkan bahwa nilai *p-value* lebih kecil dari nilai α, sehingga dapat disimpulkan bahwa terdapat perbedaan signifikan antara metode C4.5 dan AWEIG-C4.5.

Tabel 2. Spesifikasi Komputer yang Digunakan

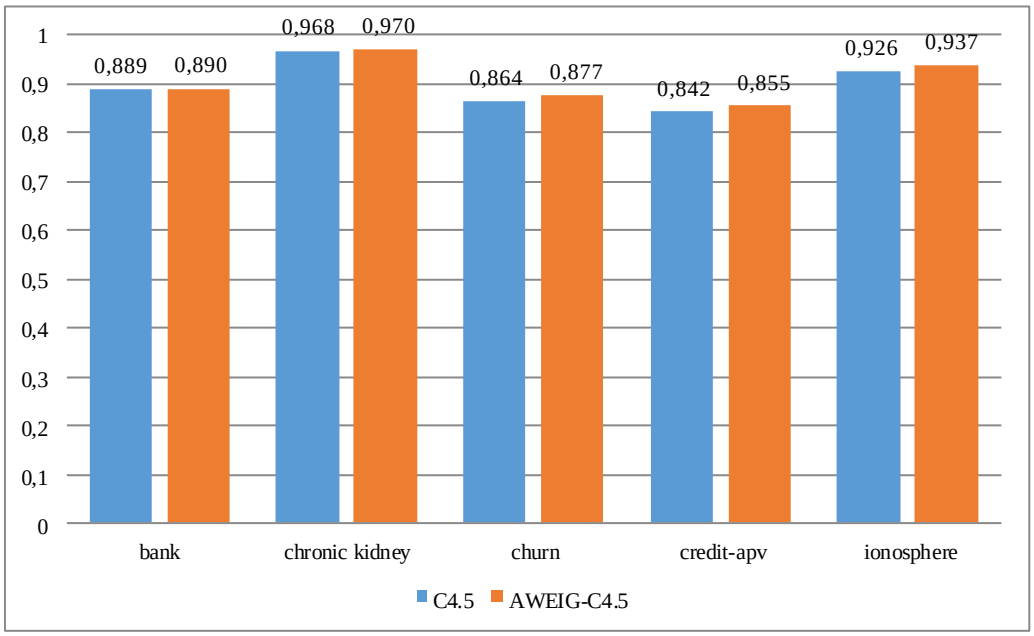
Processor	Intel ® Core ™ i5-6200 CPU @ 2.30 GHz 2.40 GHz
Memori	4 GB
Hardisk	1 TB
Sistem Operasi	Windows 10 Enterprise 64-bit
Aplikasi	Rapidminer versi 7.2.001, Microsoft Excel 2013 dan XLSTAT 2016

Tabel 3. Rekap Pengukuran Akurasi pada Metode C4.5 dan AWEIG-C4.5

No	*Dataset*	*C4.5*	*AWEIG-C4.5*
1	*Bank Marketing*	0,889	**0,890**
2	*Chronic Kidney*	0,968	**0,970**
3	*Churn*	0,864	**0,877**
4	*Credit Approval*	0,842	**0,855**
5	*Ionosphere*	0,926	**0,937**

Tabel 4. Hasil Uji *Paired t-Test* pada Metode C4.5 dan AWEIG-C4.5

	C4.5	*AWEIG-C4.5*
Mean	0,8978	**0,9058**
Variance	0,002514	0,002189
Observations	5	5
Pearson Correlation	0,994731	
Hypothesized Mean Difference	0	
df	4	
t Stat	-2,98142	
P(T<=t) one-tail	0,020341	
t Critical one-tail	2,131847	
P(T<=t) two-tail	**0,040682**	
t Critical two-tail	2,776445	



Gambar 3. Diagram Perbandingan Akurasi pada Metode C4.5 dan AWEIG-C4.5

5. Kesimpulan

Hasil penelitian menunjukkan bahwa metode AWEIG-C4.5 memberikan akurasi lebih baik daripada metode C4.5 dengan nilai rata-rata akurasi masing-masing metode adalah 0,906 dan 0,898. Uji *paired t-Test* antara metode C4.5 dan metode AWEIG-C4.5 terdapat perbedaan signifikan karena nilai *p-value* (0,04) lebih kecil daripada nilai α (0,05). Penelitian ini telah memberikan kontribusi yaitu pembobotan atribut, atribut yang dipilih adalah atribut dengan nilai bobot lebih besar daripada nilai rata-rata bobot atribut sehingga dapat mengatasi data

berdimensi tinggi, yaitu data yang mempunyai banyak atribut dan masing-masing atribut tidak relevan.

Namun ada beberapa metode yang bisa dilakukan penelitian selanjutnya agar mendapatkan hasil yang lebih baik. Metode seleksi fitur yang digunakan pada penelitian ini menggunakan metode *filter*, untuk penelitian akan datang bisa dibandingkan menggunakan metode *wrapper* atau *hybrid*. Selain itu algoritma klasifikasi yang digunakan pada penelitian ini adalah algoritma C4.5, sehingga untuk penelitian akan datang dapat dibandingkan dengan algoritma klasifikasi lainnya, seperti k-NN, Support Vector Machine (SVM), *Neural Network* (NN), dan lain-lain.

Acknowledgement

Penulis ingin mengucapkan terima kasih kepada RSW Intelligent Systems Research Group untuk diskusi yang hangat tentang penelitian ini.

Referensi

Bennasar, M., Hicks, Y., & Setchi, R. (2015). Feature selection using Joint Mutual Information Maximisation. *Expert Systems with Applications*, *42*(22), 8520–8532. http://doi.org/10.1016/j.eswa.2015.07.007

Bolón-Canedo, V., Porto-Díaz, I., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2014). A framework for cost-based feature selection. *Pattern Recognition*, *47*(7), 2481–2489. http://doi.org/10.1016/j.patcog.2014.01.008

Chen, G., & Chen, J. (2015). A novel wrapper method for feature selection and its applications. *Neurocomputing*, *159*(1), 219–226. http://doi.org/10.1016/j.neucom.2015.01.070

Chen, K. H., Wang, K. J., Wang, K. M., & Angelia, M. A. (2014). Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. *Applied Soft Computing Journal*, *24*, 773–780. http://doi.org/10.1016/j.asoc.2014.08.032

Chou, J. S., & Wu, C. C. (2013). Estimating software project effort for manufacturing firms. *Computers in Industry*, *64*(6), 732–740. http://doi.org/10.1016/j.compind.2013.04.002

Cil, I. (2012). Consumption universes based supermarket layout through association rule mining and multidimensional scaling. *Expert Systems with Applications*, *39*(10), 8611–8625. http://doi.org/10.1016/j.eswa.2012.01.192

Daliri, M. R. (2013). Chi-square distance kernel of the gaits for the diagnosis of Parkinson's disease. *Biomedical Signal Processing and Control*, *8*(1), 66–70. http://doi.org/10.1016/j.bspc.2012.04.007

Dawson, C. W. (2011). *Projects in Computing and Information Systems*. *Information Systems Journal* (Vol. 2). Retrieved from http://www.sentimentaltoday.net/National_Academy_Press/0321263553.Addison.Wesley.Publishing.Company.Projects.in.Computing.and.Information.Systems.A.Students.Guide.Jun.2005.pdf

Farquad, M. A. H., Ravi, V., & Raju, S. B. (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing Journal*, *19*, 31–40. http://doi.org/10.1016/j.asoc.2014.01.031

Fdez-Glez, J., Ruano-Ordas, D., Ramon Mendez, J., Fdez-Riverola, F., Laza, R., & Pavon, R. (2015). A dynamic model for integrating simple web spam classification techniques. *Expert Systems With Applications*, *42*(21), 7969–7978. http://doi.org/10.1016/j.eswa.2015.06.043

Gorunescu, F. (2011). *Data mining: concepts and techniques*. *Chemistry &* http://doi.org/10.1007/978-3-642-19721-5

Hajek, P., & Michalak, K. (2013). Feature selection in corporate credit rating prediction. *Knowledge-Based Systems*, *51*, 72–84. http://doi.org/10.1016/j.knosys.2013.07.008

Ispandi, I., & Wahono, R. (2015). Penerapan Algoritma Genetika untuk Optimasi Parameter pada Support Vector Machine untuk Meningkatkan Prediksi Pemasaran Langsung.

Journal of Intelligent Systems, *1*(2), 115–119. Retrieved from http://journal.ilmukomputer.org/index.php/jis/article/view/53

Jiawei Han, Micheline Kamber, J. P. (2012). *Data Mining Concepts and Techniques Third Edition*. *Elsevier and Morgan Kaufmann* (Vol. 1). http://doi.org/10.1017/CBO9781107415324.004

Jin, C., Jin, S. W., & Qin, L. N. (2012). Attribute selection method based on a hybrid BPNN and PSO algorithms. *Applied Soft Computing Journal*, *12*(8), 2147–2155. http://doi.org/10.1016/j.asoc.2012.03.015

Koprinska, I., Rana, M., & Agelidis, V. G. (2015). Correlation and instance based feature selection for electricity load forecasting. *Knowledge-Based Systems*, *82*, 29–40. http://doi.org/10.1016/j.knosys.2015.02.017

Korfiatis, V. C., Asvestas, P. A., Delibasis, K. K., & Matsopoulos, G. K. (2013). A classification system based on a new wrapper feature selection algorithm for the diagnosis of primary and secondary polycythemia. *Computers in Biology and Medicine*, *43*(12), 2118–2126. http://doi.org/10.1016/j.compbiomed.2013.09.016

Larose, D. T. (2005). *Discovering knowledge in data*. *Journal of Chemical Information and Modeling* (Vol. 53). A John Willey & Sons, Inc., Publication. http://doi.org/10.1017/CBO9781107415324.004

Moustra, M., Avraamides, M., & Christodoulou, C. (2011). Artificial neural networks for earthquake prediction using time series magnitude data or Seismic Electric Signals. *Expert Systems with Applications*, *38*(12), 15032–15039. http://doi.org/10.1016/j.eswa.2011.05.043

Oliver, A., Torrent, A., Lladó, X., Tortajada, M., Tortajada, L., Sentís, M., … Zwiggelaar, R. (2012). Automatic microcalcification and cluster detection for digital and digitised mammograms. *Knowledge-Based Systems*, *28*, 68–75. http://doi.org/10.1016/j.knosys.2011.11.021

Qian, W., & Shu, W. (2015). Mutual information criterion for feature selection from incomplete data. *Neurocomputing*, *168*, 210–220. http://doi.org/10.1016/j.neucom.2015.05.105

Sáez, J. A., Derrac, J., Luengo, J., & Herrera, F. (2014). Statistical computation of feature weighting schemes through data estimation for nearest neighbor classifiers. *Pattern Recognition*, *47*(12), 3941–3948. http://doi.org/10.1016/j.patcog.2014.06.012

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. http://doi.org/10.1145/505282.505283

Sun, X., Liu, Y., Xu, M., Chen, H., Han, J., & Wang, K. (2013). Feature selection using dynamic weights for classification. *Knowledge-Based Systems*, *37*, 541–549. http://doi.org/10.1016/j.knosys.2012.10.001

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, *41*(2), 321–330. http://doi.org/10.1016/j.eswa.2013.07.046

Verónica Bolón-Canedo, Noelia Sánchez-Maroño, A. A.-B. (2015). *Feature selection for high-dimensional industrial data. Springer International Publishing Switzerland 2015*. Springer International Publishing Switzerland 2015. http://doi.org/10.1007/s13748-015-0080-y

Wahono, R. S., Suryana, N., & Ahmad, S. (2014). Metaheuristic Optimization based Feature Selection for Software Defect Prediction. *Journal of Software Engineering*, *9*(5), 1324–1333. http://doi.org/10.4304/jsw.9.5.1324-1333

Wang, C. M., & Huang, Y. F. (2009). Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data. *Expert Systems with Applications*, *36*(3 PART 2), 5900–5908. http://doi.org/10.1016/j.eswa.2008.07.026

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). *Top 10 algorithms in data mining*. *Knowledge and Information Systems* (Vol. 14). http://doi.org/10.1007/s10115-007-0114-2

Zhao, M., Fu, C., Ji, L., Tang, K., & Zhou, M. (2011). Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes. *Expert Systems with Applications*, *38*(5), 5197–5204. http://doi.org/10.1016/j.eswa.2010.10.041