(P) ISSN 2442-4512
(O) ISSN 2503-3832

KOMPARASI ALGORITMA KLASIFIKASI DENGAN PENDEKATAN LEVEL DATA UNTUK MENANGANI DATA KELAS TIDAK SEIMBANG

Ahmad Ilham[1]

Program Studi Teknik Informatika
Universitas Dian Nuswantoro, Semarang, Jawa Tengah
dapurachmadilham@gmail.com

ABSTRAK

Masalah data kelas tidak seimbang memiliki efek buruk pada ketepatan prediksi data. Untuk menangani masalah ini, telah banyak penelitian sebelumnya menggunakan algoritma klasifikasi menangani masalah data kelas tidak seimbang. Pada penelitian ini akan menyajikan teknik under-sampling dan over-sampling untuk menangani data kelas tidak seimbang. Teknik ini akan digunakan pada tingkat preprocessing untuk menyeimbangkan kondisi kelas pada data. Hasil eksperimen menunjukkan *neural network* (NN) lebih unggul dari *decision tree* (DT), *linear regression* (LR), *naïve bayes* (NB) dan *support vector machine* (SVM).

Kata Kunci: under-sumpling, over-sumpling, data kelas tidak seimbang, algoritma klasifikasi.

ABSTRACT

The Problems unbalancing class dataset have an adverse effect on the accuracy of prediction data. To deal with this problem, it has been many previous studies using classification algorithms handle the data class is not balanced. This research will present the technique under-sampling and over-sampling to handle the data is not balanced. This technique is used at the level of the preprocessing to balance class condition on the data. The comparison showed neural network (NN) is superior decision tree (DT), linear regression (LR), naïve Bayes (NB) and support vector machine (SVM).

Keywords: under-sumpling, over-sumpling, data class unbalanced, classification algorithm.

(P) ISSN 2442-4512
(O) ISSN 2503-3832

1. PENDAHULUAN

Masalah data kelas tidak seimbang sering disebabkan oleh satu kelas kalah banyak dengan kelas lain didalam dataset [1][2]. Masalah ini banyak dijumpai diberbagai domain aplikasi seperti pada deteksi tumpahan minyak [4], pengindraan jarak jauh [5] klasifikasi teks [6], pemodelan respon [7], penilaian kualitas data sensor [8], deteksi kartu kredit palsu [9] dan extraksi pengetahuan dari database [10] sehingga hal ini menjadi penting bagi para peneliti di bidang data mining [11]. Namun dalam maslah ini cukupa sulit karena algoritma klasifikasi tradisional bias terhadap kelas minoritas [12], artinya apabila dipaksakan hasil prediksi dapat mendekati keliru bahkan salah [13].

Telah banyak penelitian yang dilakukan untuk mengatasi masalah data kelas tidak seimbang, seperti yang dilakukan oleh Zhou dan Liu [14] menunjukkan bahwa menyelesaikan masalah multi-class lebih sulit dari dua class. Penelitian lain juga menunjukkan bahwa algoritma standar tidak bekerja pada data kelas tidak seimbang atau masalah data multi class [15], sehingga menggunakan metode tradisional kurang tepat menangani masalah ini, apabila dipaksakan dapat menimbulkan prediksi bias dan hasil akurasi yang menyesatkan [16].

Solusi untuk data kelas tidak seimbang dibagi menjadi kategori level data dan level algoritma [17]. Metode pada level data mengubah distribusi dataset menjadi seimbang kemudian dipelajari untuk meningkatkan deteksi kelas minoritas. Sedangkan metode dilevel algoritma memodifikasi algoritma data mining yang diajukan untuk menyelesaikan masalah data kelas tidak seimbang. Under-sampling dan over-sampling merupakan turunan dari kategori level data, kemudian SVM, k-NN, SMOTE, *Adaptive Synthetics*, *Random Forest* adalah turunan dari level algoritma.

Over-sampling bekerja dengan kelas mayoritas, memiliki kelebihan pada dataset yang besar seperti mengurangi jumlah pengamatan dari kelas mayoritas untuk membuat kumpulan data seimbang, meningkatkan run time. Namun juga memiliki kelemahan pada kurangnya informasi penting yang berada di kelas mayoritas yang dihapus [18]. Over-sampling bekerja dengan kelas minoritas, memiliki kelebihan menyeimbangkan data dengan teknik acak, tidak meniadakan atau menghapus pengamatan, akan tetapi dengan adanya replikasi pengamatan pada data asli dapat menyebabkan *over fitting*, walaupun akurasi tinggi [19].

Pada penelitian ini metode under-sumpling dan over-sumpling digunakan ditahap preprosesing. Selanjutnya hasil dari preprosesing akan dijadikan training dan testing menggunakan validasi silang dimana didalamnya terdapat algoritma klasifikasi. Algoritma klasfikasi yang digunakan diantaranya *decicion tree* (DT), *neural network* (NN), *linear regression* (LR), *naïve bayes* (NB) dan SVM. Selanjutnya evaluasi algoritma akan menggunakan akurasi yakni *Area Under Curve* (AUC) sebagai indikator utama menentukan algoritma yang terbaik dalam pengklasifikasian data kelas tidak seimbang. Metode 10-*Cross Validation* digunakan untuk mencegah timbulnya hasil yang beragam.

Penelitian ini disusun sebagai berikut. Pada bagian 2, karya-karya istimewa dijelaskan. Pada bagian 3, metode yang diusulkan disajikan. Hasil eksperimen membandingkan metode yang diusulkan dengan orang lain disajikan pada bagian 4. Akhirnya, pekerjaan kami dari makalah ini diringkas dalam bagian terakhir.

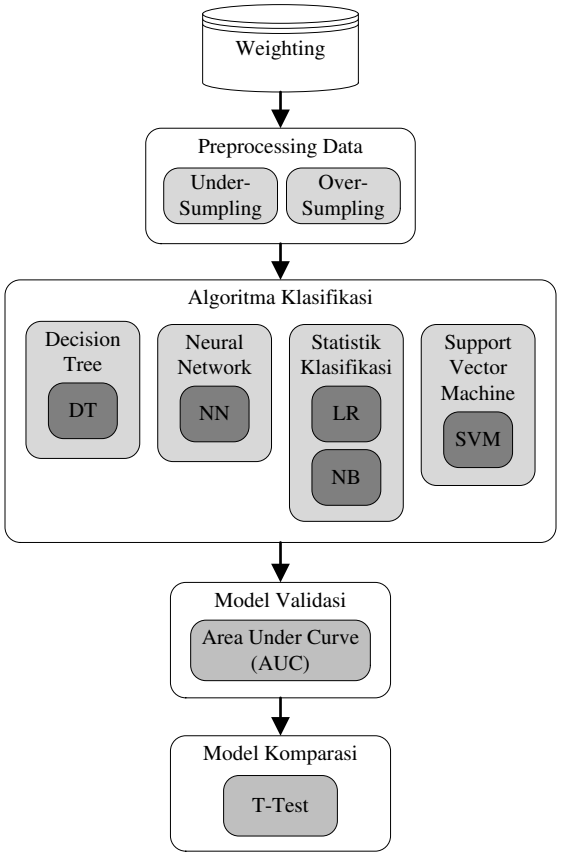
2. TINJAUAN PUSTAKA

Menangani masalah data kelas tidak seimbang merupakan tantangan cukup serius bagi para analisis data. Maka dari itu telah banyak laporan penelitian yang dilakukan untuk solusi masalah ini.

Penelitian pada didibidang pre-processing data telah banyak dilakukan antara lain, Wilson (1972) [20] mencoba untuk mengaplikasikan edited *nearest neighbour rule* (ENN) untuk mengurangi sampel pre-classified untuk penelitiannya. Lalu Tomek (1976) [21] mengusulkan *Tomek Link* untuk menentukan batas klasifikasi kelas lalu dilanjutkan oleh Gowda dan Krisna (1979) [22] mengusulkan *condensed nearest neighbor* (CNN) yang berhasil menemukan pasangan titik-titik berpasangan yang berpartisipasi dalam pembentukan batas *piecewise-linier*. Kubat dan Matwin (1997) [4] mengusulkan *one sided selection* (OSS) untuk menangani masalah data kelas tidak seimbang, dan masih banyak lagi. Dari uraian diatas dapat disimpulkan bahwa preprocessing data dalam topik penelitian ini masih relevan karena sangatlah penting untuk menemukan akurasi dimana jauh dari kesimpulan yang bias.

3. METODE YANG DIUSULKAN

Kerangka kerja yang diusulkan dapat dilihat pada Gambar 3.1. Kerangka kerja terdiri dari 1) dataset, 2) preprosesing data, 3) algoritma yang diusulkan, 4) membangun validasi, 5) membangun model evaluasi, 6) membangun model komparasi



Gambar 3.1 Kerangka kerja yang diusulkan

3.1 Dataset

Pada makalah ini dataset yang digunakan adalah dataset *public*, sehingga penelitian ini dapat diulang, diuji coba dan diverifikasi [23].

Pada makalah ini digunakan dataset '*glass4'* atribut dataset, deskripsi dan nilainya dapat dilihat pada Tabel 3.1. Dataset ini diambil dari KEEL repository.

Atribut	Tipe Data	[min-max]
RI	Real	[1.53393 , 1.51115]
Na	Real	[17.38, 10.73]
Mg	Real	[4.49, 0]
Al	Real	[3.5, 0]
Si	Real	[75.41, 0]
K	Real	[6.21, 0]
Ca	Real	[16.19, 0]
Ba	Real	[3.15, 0]
Fe	Real	[0.51, 0]
Class	Positif, Negatif	[201, 13]

Tabel 3.1 Karakteristik data *glass4*

3.2 Klasifikasi Algoritma

Pada makalah ini akan dibandingkan beberapa algoritma yang yang dikenal secara umum, antara lain *decision tree* (DT) *neural nerwork* (NN), *linear regression* (LR), *naïve bayes* (NB) dan SVM yang kemudian untuk didapatkan akurasi terbaik pada kasus data weighting.

3.3 Membangun Model Validasi

Ukuran validasi yang digunakan merujuk pada *state-of-the-art* adalah 10-fold cross validation, karena metode ini telah menjadi standar dalam penelitian praktis [24]. Metode validasi ini berarti keseluruhan data dibagi menjadi 10 bagian sama besar dan kemudian dilakukan proses *learning* sebanyak 10 kali. Pada Tabel 3.2 dapat dilihat bahwa saat salah satu bagian dijadikan data testing, maka ke sembilan bagian data lainnya akan dijadikan sebagai data learning. Setelah itu dihitung rata rata akurasi dari masing masing iterasi untuk mendapatkan akurasinya. 10-*fold cross-validation* ini sudah menjadi standard dari penelitian akhir akhir ini, dan beberapa penelitian juga didapatkan bahwa penggunaan stratifikasi dapat meningkatkan hasil yang lebih tidak beragam [24].

n-validation	Partisi Dataset									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

Tabel 3.2 Stratified 10-*fold cross-validation*

3.4 Membangun Model Komparasi

Pada makalah ini akan menggunakan akurasi popular dengan sebutan *Area Under Curve* (AUC) untuk mengukur performa akurasi algoritma. Pada umumnya algoritma yang memiliki nilai AUC diatas 0.6 mempunyai performa yang cukup efektif. Pada Table 3.3 menunjukkan interpretasi dari masing-masing nilai AUC.

AUC value	Meaning
0.90 – 0.100	Excellent classification
0.80 – 0.90	Good classification
0.70 – 0.80	Fair classification
0.60 – 0.70	Poor classification
< 0.60	Failure

Tabel 3.3 Nilai AUC dan interpretasinya

3.5 Membangun Model Komparasi

Untuk membangun model komparasi akan menggunakan test parametric dan non-parametric, dimana T-test digunakan untuk test parametrik dan friedman test digunakan untuk test non-parametrik.

4. HASIL PENELITIAN

Eksperimen dilakukan menggunakan laptop berbasis Intel Celeron 2.16 GHz CPU, 2 GB RAM

(P) ISSN 2442-4512
(O) ISSN 2503-3832

dan sistem operasi Windows 10 Professional 64-bit. Aplikasi yang digunakan adalah RapidMiner 7.2 library.

Pada makalah ini digunakan 1 dataset real '*glass4'* dari *KEEL repository*. Pada Tabel 3.1 dapat dilihat atribut atribut yang terdapat dalam dataset.

4.1 Preprocessing Data

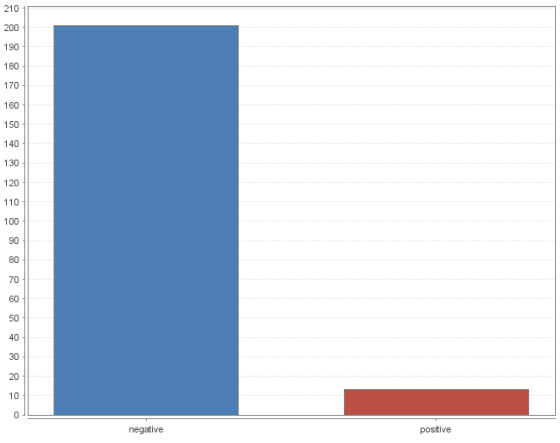
Pada Tabel 3.4 data diuji menggunakan algoritma Decision Tree, model evaluasi menggunakan 10-*fold cross validation*. Kemudian mendapat akurasi 86.40%, lalu pada bagian confusion matrix, true negative 33 dan true positive 35. Dapat kita lihat bahwa hasil proses data terjadi bias, dapat dideteksi pada gambar plot. Pada gambar plot menjelaskan bahwa kelas positih lebih condong berat dari kelas negative.

accuracy: 86.40% +/- 4.96% (mikro: 86.40%)

	true negative	true positive	class precision
pred. negative	205	35	85.42%
pred. positive	33	227	87.31%
class recall	86.13%	86.64%	

Tabel 4.1 Performance

Selanjutnya Gambar 4.1 menunjukkan terjadinya perubahan pola data yang signifikan di kelas minoritas negative berkurang dan kelas positif, berbeda dengan kelas mayoritas menempati urutan terbanyak masih yang disebabkan penghapusan informasi dari gaya teknik ditambah dengan menunjukkan perubahan akurasi dan mencapai 92.97% lalu di bagian confusion matrix terjadi perubahan, true positive 2 dan true negative 18, terjadi perubahan signifikan namun belum seutuhnya seimbang.



accuracy: 92.97% +/- 2.75% (mikro: 92.98%)

	true positive	true negative	class precision
pred. positive	260	18	93.53%
pred. negative	2	5	71.43%
class recall	99.24%	21.74%	

Gambar 4.1 Plot & Performance vector under-sampling

Pada Tabel 4.3 menunjukkan nilai performance yang seimbang. Confusion matrix menunjukkan true positive 5 dan true negative 5 namun perubahan ini juga ikut menurunkan akurasinya sehingga menjadi 85.71%. Artinya adalah metode udersampling dan oversampling memberi informasi berharga yang saling melengkapi sehingga akurasi yang diperoleh dapat dipertanggung jawabkan.

accuracy: 85.71% +/- 12.78% (mikro: 85.71%)

	true positive	true negative	class precision
pred. positive	47	5	90.38%
pred. negative	5	13	72.22%
class recall	90.38%	72.22%	

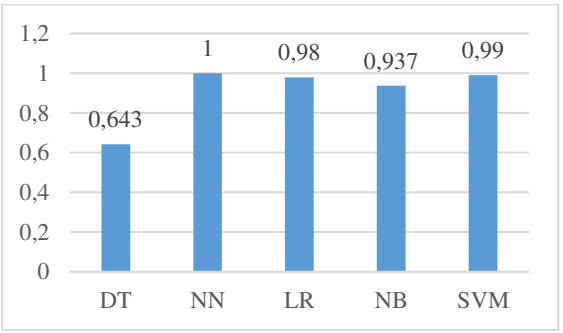
Tabel 4.2 Performance vectore over sampling

4.2 Komparasi Algoritma

Tabel 4.3, menginformasikan area under curve (AUC) dari semua algoritma klasifikasi yang digunakan dimana dapat dilihat bahwa algoritma Neural Network (NN) memiliki nilai AUC paling tinggi yaitu 1.

Algortima	DT	NN	LR	NB	SVM
AUC	0.643	1	0.98	0.937	0.99

Tabel 4.3 AUC dari semua algoritma klasifikasi

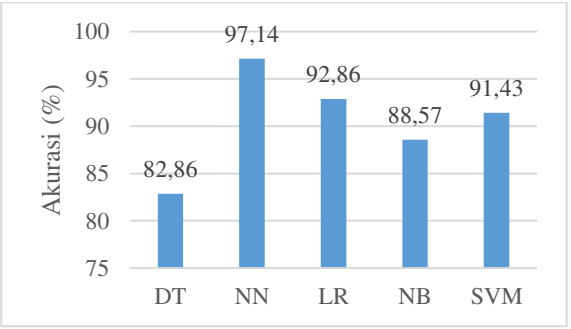


Gambar 4.4 Grafik AUC dari semua algoritma klasifikasi

Sedangkan untuk confusion matrix yang mengukur tingkat akurasi menghasilkan nilai tertinggi NN 97.14%, dapat lihat di Tabel 4.2 untuk masing masing nilai akurasi dari semua algoritma yang digunakan.

Algoritma	DT	NN	LR	NB	SVM
Confusion Matrix (%)	82.86	97.14	92.86	88.57	91.43

Tabel 4.2. Akurasi dari semua algoritma pengklasifikasi



Gambar 4.2 Grafik akurasi Confussion Matrix dari semua pengklasifikasi

Dari kedua variabel evaluasi diatas, dilakukan pengujian perbandingan antara masing-masing

variabel dengan menggunakan Pairwise T-test sehingga didapat hasil algoritma NN lebih unggul diikuti oleh LR, SVM lalu NB (lihat Gambar 4.3).

A	B	C	D	E	F
	0.829 +/- 0.125	0.971 +/- 0.057	0.929 +/- 0.096	0.886 +/- 0.107	0.914 +/- 0.095
0.829 +/- 0.125		0.004	0.059	0.285	0.100
0.971 +/- 0.057			0.240	0.038	0.120
0.929 +/- 0.096				0.358	0.741
0.886 +/- 0.107					0.535
0.914 +/- 0.095					

Gambar 4.3 Hasil T-*test*

Dari hasil *Pairwise* T-*test* dapat kita melihat apakah ada perbedaan signifikan antara masing-masing algoritma, untuk lebih jelasnya dapat dilihat pada Tabel 4.3.

Keterangan: Y = ada T = tidak	DT	NN	LN	NB	SVM
DT	T	**Y**	T	T	T
NN	T	T	T	**Y**	T
LR	T	T	T	T	T
NB	T	T	T	T	T
SVM	T	T	T	T	T

Tabel 4.3 Perbedaan signifikan dari hasil T-*tetst*

5 KESIMPULAN

Pendekatan level data dengan menggunakan teknik under-sampling dan over-sampling bertujuan menangani data kelas tidak seimbang. Selanjutnya algoritma klasifikasi yang diusulkan dikomparasi untuk membandingkan kinerja algoritma klasifikasi. Dataset yang digunakan adalah *weighting* merupakan data sintetic dari UCI repository, lima algoritma klasifikasi, model 10-*fold cross validation* dan AUC sebagai indikator akurasi. T-*test* digunakan untuk menguji perbedaan signifikan pada AUC antar model. Hasil percobaan menunjukkan NN lebih unggul pada dataset '*weighting'*. LR, SVM dan NB juga tampil baik dan secara statistic hampir tidak ada perbedaan signifikan antar algoritma pengklasifikasi.

Daftar Pustaka

[1] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, pp. 1623–1637, 2015.

[2] N. V Chawla, N. Japkowicz, and P. Drive, "Editorial : Special Issue on Learning from Imbalanced Data Sets," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 1–6, 2004.

[3] Mi. Kubat, R. Holte, and S. Matwin, "Learning when Negatif Example Abound," *Mach. Learn. ECML-97*, vol. 1, 1997.

[4] M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One Sided Selection," *Proc. Fourteenth Int. Conf. Mach. Learn.*, vol. 4, no. 1, pp. 179–186, 1997.

[5] L. Bruzzone and S. B. B. Serpico, "Classification of imbalanced remote-sensing data by neural networks," *Pattern Recognit. Lett.*, vol. 18, pp. 1323–1328, 1997.

[6] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Mach. Learn.*, vol. 30, no. 2–3, pp. 195–215, 1998.

[7] H. Shin and S. Cho, "Response modeling with support vector machines," *Expert Syst. Appl.*, vol. 30, no. 4, pp. 746–760, 2006.

[8] A. Rahman, D. V. Smith, and G. Timms, "Multiple classifier system for automated quality assessment of marine sensor data," *2013 IEEE Eighth Int. Conf. Intell. Sensors, Sens. Networks Inf. Process.*, pp. 362–367, 2013.

[9] A. Agrawal, H. L. Viktor, and E. Paquet, "SCUT : Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Undersampling," vol. 1, no. Ic3k, pp. 226–234, 2015.

[10] A. Bhardwaj, A. Tiwari, H. Bhardwaj, and A. Bhardwaj, "A Genetically Optimized Neural Network Model for Multi-class Classificatio," *Expert Syst. Appl.*, 2016.

[11] X. Wu, V. Kumar, Q. J. Ross, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.

[12] S. García and F. Herrera, "Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy.," *Evol. Comput.*, vol. 17, no. 3, pp. 275–306, 2009.

[13] Y. Tang, Y. Q. Zhang, and N. V. Chawla, "SVMs modeling for highly imbalanced classification," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 39, no. 1, pp. 281–288, 2009.

[14] L. Abdi and S. Hashemi, "To Combat Multi-Class Imbalanced Problems by Means of Over-Sampling Techniques," vol. 28, no. 1, pp. 238–251, 2016.

[15] G. Menardi and N. Torelli, *Training and assessing classification rules with imbalanced data*, vol. 28, no. 1. 2014.

[16] A. Fernandez, V. Lopez, M. Galar, M. J. Del Jesus, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowledge-Based Syst.*, vol. 42, pp. 97–110, 2013.

[17] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *Adv. Intell. Comput.*, vol. 17, no. 12, pp. 878–887, 2005.

[18] C. Drummond and R. C. Holte, "C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," *Work. Learn. from Imbalanced Datasets II*, pp. 1–8, 2003.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16,

(P) ISSN 2442-4512
(O) ISSN 2503-3832

no. January, pp. 321–357, 2002.

[20] D. L. Wilson, "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data," *IEEE Trans. Syst. Man Cybern.*, vol. 2, no. 3, pp. 408–421, 1972.

[21] I. Tomek, "Two Modification of CNN," pp. 769–772, 1976.

[22] K. C. Gowda and G. Krishna, "The Condensed Nearest Neighbor Rule Using the Concept of Mutual Nearest Neighborhood," *IEEE Trans. Inf. Theory*, vol. 25, no. 4, pp. 488–490, 1979.

[23] C. Catal and B. Diri, "Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem," *Inf. Sci. (Ny).*, vol. 179, no. 8, pp. 1040–1058, 2009.

[24] I. H. Witten, E. Frank, and M. a. Hall, *Data Mining Practical Machine Learning Tools and Techniques Third Edition*, vol. 277, no. Tentang Data Mining. 2011.