ANALISIS KOMPARASI ALGORITMA KLASIFIKASI UNTUK MENANGANI DATA TIDAK SEIMBANG PADA DATA KEBAKARAN HUTAN

Castaka Agus Sugianto
Program Studi Teknik Informatika, Politeknik TEDC Bandung, 40513 Indonesia
E-mail : castaka@poltektedc.ac.id

Abstrak

Untuk menghasilkan hasil yang maksimal di dalam proses klasifikasi data harus memiliki distribusi yang sama dengan data pelatihan. Namun, kenyataannya data seperti ini, tidak selalu ditemukan banyak juga data yang distribusinya tidak sama, dimana satu kelas mungkin diwakili oleh data dengan jumlah yang besar, sementara kelas yang lain diwakili oleh hanya beberapa. Algoritma klasifikasi data mining banyak yang dapat digunakan untuk menangani data tidak seimbang, maka dari itu perlu dilakukan komparasi untuk mengetahui seberapa tinggi tingkat akurasi dari masing-masing algoritma yang ada. Algoritma yang digunakan adalah K-Means + C4.5, K-Means + Naïve Bayes, K-Means + Random Forest dan K-Means + Neural Network. Dataset terdiri dari dua kombinasi, yang terdiri dari variabel meteorologi dan fire weather index (FWI) untuk memprediksi ukuran kebakaran hutan. Hasil dari proses klasifikasi dievaluasi dengan menggunakan cross validation, confusion matrix, dan T-Test

Kata Kunci : *Algoritma C4.5, Naïve Bayes, Random Forest, Neural Network, K-Means, Data tidak seimbang, Data Mining.*

Abstract

To produce maximum results in the process of data classification should have the same distribution with the training data. However, the fact is this kind of data, not always be found many data distribution is not balanced, where one class may be represented by the majority of data, while the other class is represented by the minority class. Classification of data mining algorithms much can be used to handle the imbalanced dataset, therefore it is necessary for comparing to determine how high the level of accuracy of each algorithm available. K-Means + C4.5, K-Means + Naïve Bayes, K-Means + Random Forest and K-Means + Neural Network Algorithms are implemented. The dataset is composed of two combinations, consisting of variable meteorological and fire weather index (FWI) to predict the size of forest fires. The results from the classification process was evaluated using cross validation, confusion matrix, and T-Test.

Keywords : *Algorithm C4.5, Naïve Bayes, Random Forest, Neural Network K-Means, imbalanced dataset, Data Mining.*

1. PENDAHULUAN

Data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola dan hubungan dalam set data berukuran besar [1]. Algoritma 10 terartas didalam data mining yaitu : C4.5, K-Means, SVM, Apriori, EM, PageRenk, AdaBoost, kNN, Naive bayes, and CART [2]. Algoritma tersebut sudah banyak diterapkan diberbagai domain dan sukses menghasilkan akurasi yang maksimal. Metode klasifikasi merupakan bagian penting dari data mining, algoritma klasifikasi beroprasi pada data yang jumlah distribusi kelasnya sama.

Namun, ada juga data yang memiliki jumlah kelas yang tidak seimbang antar kelas yang satu dengan yang lainya disebut "*imbalaced data sets*". Di dalam mesin learning jika mengunakan pendekatan klasifikasi yang setandar, data tidak seimbang menghasilkan *performace* yang kurang bagus [3]. Kinerjanya yang kurang bagus karena klasifikasi standar mungkin mengabaikan pentingnya kelas minoritas karena perwakilannya dalam dataset tidak cukup kuat [4][5].

Menilai kinerja metode dan pemilihan model telah menjadi isu penting, para peneliti banyak yang melakukan komparasi algoritma klasifikasi[6][7].

Saat ini dalam prakteknya, membandingkan berbagai metode pemodelan tidak mudah Karena kriteria kinerja yang berbeda dan prosedur validasi[8]. Penampilan model juga dipengaruhi dari keseimbangan kelas dimodel klasifikasi. Prediksi yang akurat biasanya berhubungan dengan kelas minoritas, kelas minoritas biasanya memiliki hal penting yang lebih besar [9]. Salah satu cara menyelesaikan permasalahan ketidak seimbangan kelas yaitu dengan memodifikasi data trening dengan metode *oversampling* untuk kelas minoritas atau *under-sampling* untuk kelas mayoritas [9].

Berdasarkan permasalahan tersebut dipandang perlu untuk melakukan komparasi algoritma klasifikasi untuk menangani dataset yang tidak seimbang, sehingga didapat model yang sesuai dengan data yang di gunakan. peneliti coba menggunakan teknik *clustering* dalam hal ini algoritma K-Means yang digunakan untuk menangani dataset tidak seimbang. Kombinasi ini sebagai bagian dari kontribusi penelitian. Tujuan dari penelitian ini adalah melakukan analisis komparasi 4 algoritma klasifikasi data mining yaitu k-means + C4.5, K-means + Random forest, K-means + naïve bayes dan K-means + neural network, sehingga dapat diketahui model algoritma yang paling akurat dan memiliki akurasi yang tinggi untuk kelasifikasi data kebakaran hutan.

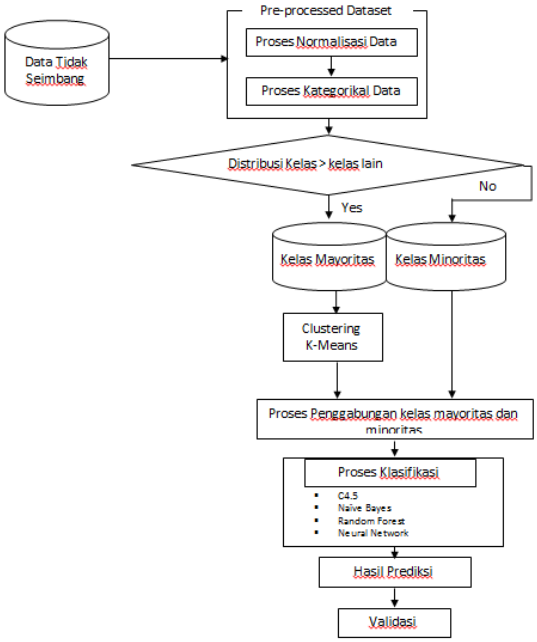
2. METODE

2.1. *Tipe Metode Penelitian*

Penelitian ini merupakan penelitian eksperimen, karena responden ditugaskan untuk berkelompok berdasarkan beberapa kriteria, yang sering disebut perlakuan variabel atau perlakuan kondisi [10].

2.2. *Metode yang diusulkan*

Skema dan pemodelan penelitian disajikan pada Gambar 1. Seluruh tahapan sebagai berikut: dimulai dengan pemilihan dataset. Pada tahap kedua, data dilakukan *preprocessing*. Pada tahap ketiga data yang diolah di *cluster* menggunakan algoritma K-Means. Terakhir data hasil *clustering* di gabungkan dengan kelas minoritas, dan kemudian dilakukan klasifikasi dengan menggunakan beberapa algoritma yaitu: C4.5, Naïve Bayes, Random Forest dan Neural Network.



Gambar 1. Metode yang di usulkan.

Seleksi data: Proses memilih data yang akan digunakan dalam proses prediksi, dataset ini dari UCI *dataset repository.*
Preprocessing Datasets: Setelah data dipilih dan kemudian dibagi menjadi tiga kategori dengan menggunakan aturan yang diperoleh kategori berikut: kecil, menengah dan besar yang mengacu pada nilai normalisasi.

Tabel 1. Area kebakaran dan kelas kebakaran hutan.

No	Area Kebakaran	Kelas kebakaran hutan
1	0– 0.99	*Small*
2	1 – 1.99	*Medium*
3	≥ 2	*Large*

Aturan kategorisasi diadopsi dari aturan penelitian yang di tulis oleh Harrison dan kawan kawan [11] dapat dilihat pada rumus nomer 1:

If *normalized (x) <1* **Then** *it is small*
If *1 ≤ normalized (x) < 2* **Then** *it is medium*
If *normalized (x) ≥ 2* **Then** *it is large* (1)

Rumus untuk menghitung normalisasi bisa dilihat pada rumus nomer 2 di bawah ini.

$$\text{Normalized}(x_i) = \frac{x_i - \bar{x}}{\delta} \quad (2)$$

*Clustering:*Untuk mengatasi masalah *dataset* tidak seimbang antara data kecil, menengah dan besar yang pertama dilakukan untuk mengurangi ukuran *dataset* kategori kecil tanpa kehilangan karakter penting dari sebuah data. Banyak metode untuk *clustering* tetapi Penelitian ini mengusulkan metode, algoritma k-means untuk memecahkan masalah ini.
Klasifikasi: Sampel diklasifikasikan ke dalam 15 kelompok yang berbeda. Hasil klaster kemudian digunakan sebagai masukan bagi *Classifier* tersebut. Klasifikasi merupakan salah satu teknik dalam data mining yang sering disebut *supervise learning*, yang digunakan untuk memisahkan data menjadi beberapa segmen. Dalam penelitian ini membandingkan algoritma C4.5, Naïve Bayes, Random Forest dan Neural Network untuk klasifikasi data tersebut.
Hasil Prediksi: Adalah output setelah proses klasifikasi.
Validasi: Pengujian dilakukan dengan menggunakan k-fold teknik validasi silang (*cross validation*). Metode *cross-validation* digunakan untuk menghindar
i
tumpang tindih dalam data pengujian.

3. Penelitian ini untuk melakukan komparasi hasil pengujian klasifikasi menggunakan *confusion matrix*. *Confusion matrix* adalah visualisasi untuk mengevaluasi model klasifikasi [12]. *Confusion matriks* berisi informasi tentang kelas yang sebenarnya dan kelas diprediksi. Bagian kolom mewakili kelas prediksi dan baris mewakili kelas yang sebenarnya. *Confusion matrix* dapat di lihat pada tabel 2 dibawah ini[13].

Tabel 2. *Confusion Matrix* dua kelas

	Predicted Class	
	Class=Small	*Class=Large*

Actual Class	*Class=Small*	*A (True Positive)*	*B (False Negative)*
	Class=Large	*C (False Positive)*	*D (True Negative)*

2.3. *Data Sample*

Data kebakaran hutan dikumpulkan dari studi Cortez dan Morais [14]] yang dapat di *download* di *UCI Machine Learning Repository*: Data Sets (http://archive.ics.uci.edu/ml/datasets/Forest+Fires). *Dataset* berisi kebakaran hutan, *forest fire weather index* (FWI) komponen dalam Montesinho Natural Park, daerah sebelah timur laut dari Portugal. Pengamatan cuaca dikumpulkan oleh Braganca Polytechnic Institute dan terintegrasi dengan dataset kebakaran hutan. Taman ini dibagi menjadi 81 lokasi yang berbeda dengan ukuran peta 9 × 9 kotak. *Dataset* memiliki total 517 sampel, dari tahun 2000 sampai tahun 2007. Adapun atribut – atributnya dapat dilihat pada tabel 3 dibawah ini.

Tabel 3. Atribut *Dataset*

Atribut	Deskripsi
X	*X - axis coordinate (from 1 to 9)*
Y	*Y - axis coordinate (from 1 to 9)*
Month	*Month of the year (January to December)*
Day	*Day (of the week (Monday to Sunday)*
FFMC	*Fine Fuel Moisture Code*
DMC	*Duff Moisture Code*
DC	*Drought Code*
ISI	*Initial Spread Index*
Temperature	*Outside temperature (in oC)*
Relative Humidity	*Outside relative humidity (in %)*
Wind	*Outside wind speed (in km/h)*
Rain	*Outside rain (in mm/m2)*
Area	*Total burned area (in ha)*
Normalized Burnt Area	*Total burned area after normalized (in ha)*
Burnt	*Transformation from normalized burnt area (Small, Medium, and Large).*

3. HASIL DAN PEMBAHASAN

3.1. *Proses data sebelum digunakan*

Total dataset yang digunakan dalam penelitian 517 sampel, data dikumpulkan dari tahun 2000 sampai tahun 2007. Pertama dataset, variabel area yang terbakar dirubah dari nilai kontinyu jadi variabel kategori. Setelah transformasi dari nilai kontinyu ke bentuk kategori, ditemukan bahwa sampel untuk kategori kecil adalah 502, sementara ada 6 sampel menengah dan 9 sampel besar. Dari distribusi tersebut sampel kecil lebih banyak dari kategori lainnya, data yang tidak seimbang memiliki efek pada kinerja metode klasifikasi. Oleh karena itu, teknik *clustering* adalah solusi untuk mengatasi efek dari data tidak seimbang.

Pada penelitian ini penulis menggunakan algoritma K-Means untuk mengatasi data yang tidak seimbang. Dalam proses *clustering*, hasil percobaan menunjukkan bahwa jumlah *cluster* terbaik adalah 13, dimana 13 jadi kelompok kecil yaitu *small_0 - small_12*. Dan kemudian cluster digabungkan dengan cluster menengah dan besar. Hasil gabungan *cluster* akan dijadikan sebagai masukan dalam proses klasifikasi.

3.2 *Proses Normalisasi data*

Normalisasi dilakukan dalam penelitian ini menggunakan rumus normalisasi, hasilnya dengan rata-rata 12,84729 dan standar deviasi 63,65582 diperoleh dari data keseluruhan, Perhitungan normalisasi data adalah sebagai berikut.

Normalisasi $(x_i) = \frac{x_i - \bar{x}}{\delta}$

Normalisasi$(x_1) = \frac{64.1 - 12.84729}{63.65582} = \frac{51.25271}{63.65582}$

=0.805154

Normalisasi$(x_2) = \frac{71.3 - 12.84729}{63.65582} = \frac{58.45271}{63.65582}$

=0.918262

Normalisasi$(x_3) = \frac{88.49 - 12.84729}{63.65582} = \frac{75.64271}{63.65582}$

=1.188308

Normalisasi$(x_4) = \frac{95.18 - 12.84729}{63.65582} = \frac{82.33271}{63.65582}$

=1.293404

Normalisasi$(x_5) = \frac{103.39 - 12.84729}{63.65582} = \frac{90.54271}{63.65582}$

=1.422379

Normalisasi$(x_{517}) = \frac{10.13 - 12.84729}{63.65582} = \frac{-2.71729}{63.65582}$

= -0.04269

3.3 *Data kategori*

Daerah kebakaran dirubah dari nilai *continuous* ke bentuk kategori, dimana variabel kategori terdiri dari kategori kecil, menengah, dan besar. Data kategori didapat dari data hasil perhitungan normalisasi. Setelah dilakukan kategori data dapat ditunjukkan pada Tabel 4. di bawah ini.

Tabel 4. Data Kebakaran Hutan Setelah Kategori.

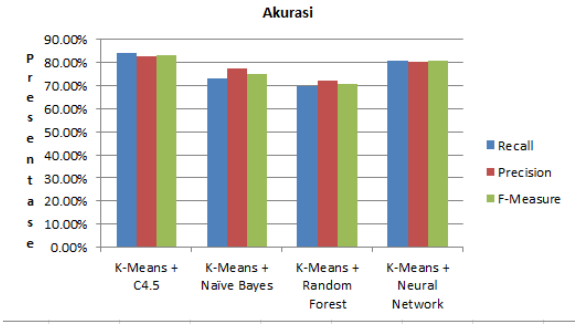
No	Normal	*Burnt*
1	0.805154	*small*
2	0.918262	*small*
3	1.188308	*medium*
4	1.293404	*medium*
5	1.422379	*medium*
6	1.45804	*medium*
7	2.23126	*large*
8	2.884775	*large*
9	2.954839	*large*
10	3.14241	*large*
…	…	...
517	-0.04269	*small*

3.4 *Hasil Perbandingan*

Penelitian ini membandingkan nilai *recall, precision* dan *F-Measure* dari hasil percobaan. Percobaan pada data tidak seimbang dalam penelitian ini membandingkan beberapa teknik klasifikasi seperti yang ditunjukkan pada Tabel 5 di bawah ini.

Tabel 5. Hasil Perbandingan *Recall, Precision* dan *F-Measure*

Algoritma	*Recall*	*Precision*	*F-Measure*
K-Means + C4.5	83.96%	82.76%	83.36%
K-Means + Naïve Bayes	73.08%	77.21%	75.09%
K-Means + Random Forest	69.60%	72.31%	70.93%
K-Means + Neural Network	80.94%	80.17%	80.55%



Gambar 2. Grafik hasil perbandingan *recall, precision dan f-measure.*

Dalam Penelitian ini pengujian yang dilakukan selain membandingkan nilai *Recall*, *Precision* dan *f-measure* juga membandingakan nilai akurasi, yang bisa dilihat pada tabel 6 di bawah ini.

Tabel 6. Hasil Perbandingan *Akurasi*

Algoritma	K-Means + Naïve Bayes	K-Means + Random Forest	K-Means + C4.5	K-Means + Neural Network
Akurasi	80.49%	79.90%	94.01%	90.12%

3.5 *Hasil Pengujian T-Test*

Untuk penentuan lebih lanjut akan digunakan pengujian dengan memanfaatkan uji statistik yaitu dengan menggunakan uji T-Test.

Tabel 7. Hasil uji T-Test

	K-Means + C4.5	K-Means + Naïve Bayes	K-Means + Random Forest	K-Means + Neural Network
K-Means		0.022	0.011	0.307

+ C4.5				
K-Means + Naïve Bayes	0.039		0.285	0.088
K-Means + Random Forest	0.011	0.144		0.074
K-Means + Neural Network	0.558	0.225	0.036	

Berdasarkan uji statistik T-Test pada tabel 7 di atas dapat dianalisis bahwa algoritma K-Means+C4.5 terhadap K-Means + Naïve Bayes ada perbedaan karena memiliki nilai Probabilitas (0,022<0,05), K-Means+C4.5 terhadap K-Means + Random Forest ada perbedaan karena memiliki nilai Probabilitas (0,011<0,05), K-Means+C4.5 terhadap K-Means + Neural Network tidak ada perbedaan karena memiliki nilai Probabilitas (0,307>0,05). Dari analisis tabel 7, bisa disimpulkan hasil pengujian nilai akurasi, *recall, precision* dan *f-measure* tidak ada perbedaan antara penggabungan algoritma K-Means+C4.5 dengan algoritma K-Means+Neural Network. Sedangkan algoritma K-Means+Neural Network hanya ada perbedaan dengan algoritma K-Means+Random Forest. Hal ini menunjukan bahwa dalam penelitian ini penggabungan algoritma K-Means+C4.5 merupakan model yang sesuai untuk menagani data ttidak seimbang pada data kebakaran hutan.

4. KESIMPULAN

Dari hasil percobaan terakhir terbukti bahwa proses *clustering* menggunakan algoritma K-Means menunjukan hasil pengujian *Recall* 83,96%, *Precision* 82,76% dan *f-measure* 83.36%. Adapun hasil menggunakan teknik k-means+Naïve Bayes dengan hasil *Recall* 73,08%, *Precision* 77,21% dan *f-measure* 75.09%. Algoritma K-means + Random Forest hasilnya *Recall* 69,60%, *Precision* 72,31% dan *f-measure* 70.93%. Sedangkan algoritma K-means + Neural Network hasilnya *Recall* 80,94%, *Precision* 80,17% dan *f-measure* 80.55%. Dari hasil pengujian akurasi penggabungan algoritma k-means + C4.5 = 94.01%, K-means + Naïve Bayes =80.49%, K-means + Random Forest =79.90% dan K-means + Neural Network = 90.12%. Hasil pengujian menunjukan bahwa Algoritma K-Means + C4.5 dapat menangani *dataset* tidak seimbang. Hal ini dapat dilihat bahwa nilai *recall* dan *precison* lebih tinggi dari nilai *recall* dan *precision* dengan menggunakan algoritma klasifikasi yang lain. Teknik *clustering* mampu menangani dataset tidak seimbang dengan membagi kelas mayoritas menjadi beberapa kelas yang lebih kecil.

DAFTAR PUSTAKA

[1] B. Santoso, *Data Mining, Teknik pemanfaatan Data untuk Keperluan Bisnis*. Jogjakarta: Geraha Ilmu, 2007.

[2] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, and others, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[3] Z. Sheng and S. Xiuyu, "Optimizing the Classification Accuracy of Imbalanced Dataset Based on SVM," in *Computer*

Application and System Modeling, 2010, vol. 0, no. Iccasm, pp. 338–341.

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

[5] C. . KrishnaVeni and T. S. Rani, "On the Classification of Imbalanced Datasets," *Computer Sciences and Technology*, vol. 2, pp. 145–148, 2011.

[6] M. Anyanwu and S. G. Shiva, "Comparative analysis of serial decision tree classification algorithms," *Journal of Computer Science and Security*, vol. 3, no. 3, pp. 230–240, 2009.

[7] B. Lee, "A comparison of support vector machines and artificial neural networks for mid-term load forecasting," *2012 IEEE International Conference on Industrial Technology*, pp. 95–101, Mar. 2012.

[8] E. Duman, Y. Ekinci, and A. Tanrıverdi, "Comparing alternative classifiers for database marketing: The case of imbalanced datasets," *Expert Systems with Applications*, vol. 39, no. 1, pp. 48–53, Jan. 2012.

[9] W. Liu, S. Chawla, and D. Cieslak, "A robust decision tree algorithm for imbalanced data sets," *Conference on Data Mining*, pp. 1–12, 2010.

[10] N. J. Salkind, *Exploring Research*, 7th ed. New Jersey: Pearson International Edition, 2009, pp. 225–230.

[11] Y. P. Yu, R. Omar, R. D. Harrison, M. K. Sammathuria, and A. R. Nik, "Pattern clustering of forest fires based on meteorological variables and its classification using hybrid data mining methods," *Computational Biology and Bioinformatics Research*, vol. 3, no. July, pp. 47–52, 2011.

[12] S. S. Imas and H. ismail Mohd, "Hotspot Occurrences Classification using Decision Tree Method," in *ICT and Knowledge Engineering*, 2010, pp. 46–50.

[13] F. Gorunescu, *Data Mining Concept Model Technique*. Springer, 2011, pp. 1–370.

[14] P. Cortez, "A data mining approach to predict forest fires using meteorological data," *Information Systems*, pp. 1 – 12, 2007.