Prosiding
ANNUAL RESEARCH SEMINAR 2016
6 Desember 2016, Vol 2 No. 1

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

Penerapan *Ensemble Stacking* Untuk Klasifikasi Multi Kelas

Rio Ananda Fitriansyah
Fakultas Ilmu Komputer, Universitas Sriwijaya
Palembang, Indonesia
rio.fitriansyah@gmail.com

Saparudin
Fakultas Ilmu Komputer, Universitas Sriwijaya
Palembang, Indonesia
saparudinmasyarif@gmail.com

***Abstrak*—Klasifikasi adalah salah satu topik utama yang banyak digunakan dalam penelitian pembelajaran mesin. Beberapa penelitian terdahulu telah menghasilkan base classifier yang sampai saat ini masih digunakan. Banyak base classifier menunjukkan performa yang baik untuk klasifikasi biner tetapi performa classifier tersebut menurun pada saat digunakan untuk klasifikasi multi-kelas. Pada penelitian sebelumnya digunakan hybrid classifier untuk klasifikasi multi kelas. Hasil penelitian menunjukkan akurasi hybrid classifier yang diajukan lebih baik dari base classifier. pada penelitian ini ensemble method stacking diterapkan. Decision tree dan naïve bayes digunakan sebagai classifier dasar. Hasil pengujian menunjukkan metode ensemble stacking hanya mampu melampui pada beberapa dataset jika dibandingkan dengan hybrid classifier.**

Kata kunci—klasifikasi, stacking, metode ensemble, metode hybrid, multi-class

I. PENDAHULUAN

Klasifikasi dalam data mining bertujuan memprediksi dan membedakan kelas data dari suatu baris data dengan atribut data tersebut diketahui namun kelasnya tidak diketahui [1]. Cara untuk memprediksi dan membedakan kelas data dengan membentuk model dari proses pembelajaran oleh *base classifier* terhadap data pelatihan. Kemudian model melakukan prediksi kelas data terhadap data ujicoba. Beberapa *base classifier* dalam data mining dapat menggunakan algoritma *Bayesian Network*, *Neural Network*, *Decision tree*, dan *Support vector machine*. Berbagai penelitian telah dilakukan untuk menyelesaikan permasalahan klasifikasi dalam dunia nyata menggunakan *base classifier* [2], [3]. Perkembangan penelitian selanjutnya bergerak ke arah penggabungan *base classifier*. Dua metode yang diajukan adalah metode *hybrid* dan *ensemble classifier*. Metode *hybrid classifier* adalah penggabungan dari paling sedikit dua teknik untuk meningkatkan performa dari teknik baru yang dihasilkan [4]. Tujuannya adalah untuk mendapatkan kelebihan-kelebihan dari beberapa teknik yang digabungkan sehingga *hybrid classifier* yang diciptakan memiliki akurasi yang lebih baik. Metode *ensemble classifier* adalah penggunaan beberapa *base classifier* secara bersamaan dalam proses klasifikasi dengan tujuan untuk meningkatkan akurasi prediksi [5]. Penelitian sebelumnya yang dilakukan oleh [6] mengajukan *hybrid classifier decision tree* dan *naïve bayes* untuk klasifikasi data multi kelas. Data *preprocessing* dilakukan untuk mereduksi jumlah baris dataset pelatihan yang dianggap dapat menurunkan tingkat akurasi prediksi. Hal ini menyebabkan dataset kehilangan sebagian informasi yang mungkin berguna. Hasil penelitian menunjukkan akurasi *hybrid classifier* yang diajukan lebih baik dari *base classifier* tunggal untuk klasifikasi multi kelas. Sementara pada penelitian yang lain [7] menerapkan metode *ensemble* pada klasifikasi biner. Perbandingan antara *ensemble classifier* terhadap *base classifier* tunggal untuk klasifikasi biner juga menunjukkan nilai akurasi yang lebih baik. Metode ensemble melakukan data *preprocessing* tanpa mereduksi baris data. Pada penelitian ini akan diterapkan metode *ensemble stacking* dengan *base classifier* yang digunakan *decision tree* (C4.5 [8]) dan *naïve bayes* [9] untuk klasifikasi multi kelas. Akurasi hasil prediksi akan dibandingkan dengan hasil prediksi metode *hybrid* dan *base classifier*.

II. TINJAUAN PUSTAKA

Pada bagian ini beberapa penelitian metode *hybrid* dan metode *ensemble* yang berkaitan dengan penelitian disajikan.

2.1. Metode Hybrid

[6] mengajukan *hybrid classifier decision tree* (C4.5) dan *naïve bayes* yang dititik beratkan pada *preprocessing* dataset pelatihan untuk menghindari *overfitting* dari pohon yang

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

dihasilkan dan tingkat akurasi mungkin akan menurun. Hasil penelitian menunjukkan akurasi kedua *hybrid classifier* dapat melampaui *base classifier*. [10] mengajukan kombinasi *classifier* C4.5 dengan metode *one against all*. Pengujian dilakukan pada tiga dataset multi kelas. Dari segi akurasi *hybrid classifier* yang diajukan lebih baik jika dibandingkan dengan *classifier* C4.5. [4] mengajukan *hybrid classifier* berdasarkan teknik *artificial bee colony optimization* dan *feature selection* algoritma *differential evolution*. Algoritma *hybrid* ini membuat subset feature yang akan digunakan untuk mereduksi dimensi data. Akurasi *hybrid artificial bee colony* dengan *feature selection differential evolution* lebih baik dari classifier lain yang dijadikan perbandingan.

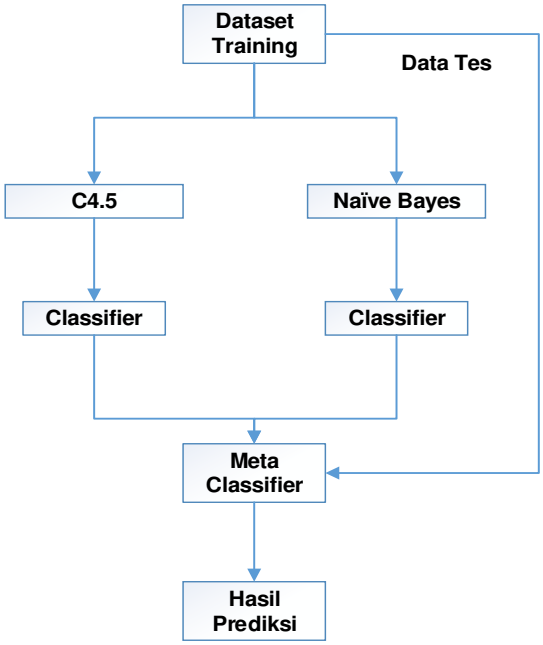
2.2. Metode Ensemble

[11] mengajukan metode *ensemble classifier selection based on clustering*. Hasil penelitian menunjukkan akurasi *ensemble classifier selection based on clustering* mampu melampaui classifier pembanding. [12] menggunakan metode *imprecise probabilities* dan *uncertainty measures* pada *decision tree* untuk menangani data dengan nilai yang hilang serta data dengan nilai kontinu untuk selanjutnya diterapkan metode *ensemble* terhadap *decision tree* yang dihasilkan.dari dua pendekatan yang digunakan pada penelitian tersebut, hasil menunjukkan kedua pendekatan menghasilkan nilai yang berbeda untuk masing-masing classifier yang digunakan. [13] mengajukan *ant colony optimization*. Menggunakan metode *stacking* [14]. Proses pembentukan model *ant colony optimization* menggunakan beberapa kandidat *base classifier* dan beberapa kandidat meta *classifier*. Akurasi *ant colony optimization* lebih baik dari classifier pembanding yang digunakan.

III. METODOLOGI PENELITIAN

Tahapan sistematis pada penelitian ini adalah pengumpulan dataset, pembelajaran dataset pelatihan, pengklasifikasian dataset ujicoba dan tahap perbandingan akurasi hasil prediksi. Dataset yang digunakan adalah dataset multikelas.

Metode *ensemble stacking* menggunakan beberapa *base classifier* pada proses pembelajaran. Ada dua tahap dalam pembelajaran *stacking*. Tahap 1, setiap *base classifier* yang digunakan dilatih dengan menggunakan dataset yang sama sehingga menghasilkan hasil prediksi masing-masing. Tahap 2, meta *classifier* mengambil hasil prediksi dari *base classifier* sebagai input untuk menentukan kelas mana yang paling mungkin terhadap data ujicoba. Diagram metode *ensemble stacking* yang digunakan adalah sebagai berikut :



Gambar 1. Metode *Stacking* C4.5 dan *Naïve bayes*

IV. HASIL DAN PEMBAHASAN

Pada bagian ini akan disajikan hasil pengujian dataset yang digunakan dengan *base classifier*, *hybrid classifier* dan *ensemble stacking classifier*.

4.1. Dataset & Perangkat Penunjang

Penelitian ini menggunakan 9 dataset dari UCI Repository. Penggunaan dataset ini ditujukan untuk menyamakan lingkungan penelitian dengan penelitian sebelumnya. 9 dataset yang digunakan adalah sebagai berikut :

1. Dataset Kanker Payudara (Breast cancer).
2. Dataset Kontak Lensa (Contact lenses).
3. Dataset Diabetes Pima Indian (Diabetes).
4. Dataset Kaca (Glass).
5. Dataset Tanaman Iris (Iris plants).
6. Dataset Tanaman Kedelai (Soybean).
7. Dataset Voting Kongres Amerika Serikat Tahun 1984 (Vote).
8. Dataset Segmentasi Gambar (Image seg.)
9. Dataset Permainan Tic-Tac-Toe (Tic-Tac-Toe).

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

Tabel 1. Dataset yang digunakan

Dataset	Jumlah Atribut	Tipe Atribut	Jumlah Baris Data	Jumlah Kelas
Breast cancer	9	Nominal	286	2
Contact lenses	4	Nominal	24	3
Diabetes	8	Real	768	2
Glass	9	Real	214	7
Iris plants	4	Real	150	3
Soybean	35	Nominal	683	19
Vote	16	Nominal	435	2
Image seg.	19	Real	1500	7
Tic-Tac-Toe	9	Nominal	958	2

4.2. Metode Pengukuran Hasil

Empat kriteria pengukuran yaitu akurasi, presisi, sensitifitas dan spesifisitas dengan pengujian 10 fold cross validation.

Persamaan kriteria pengukuran adalah sebagai berikut :

$$Akurasi = \frac{BP + BN}{BP + BN + SP + SN} \quad \text{.....1}$$

$$Presisi = \frac{BP}{BP + SP} \quad \text{.....2}$$

$$Sensitifitas = \frac{BP}{BP + SN} \quad \text{.....3}$$

$$Spesifisitas = \frac{BN}{BN + SP} \quad \text{.....4}$$

dengan BP, BN, SP dan SN melambangkan benar positif, benar negatif, salah positif dan salah negatif.

Tabel 2. Simbol dan deskripsi yang digunakan di persamaan 1-4

Simbol		Deskripsi
BP	:	Data diprediksi anggota kelas tersebut dan memang anggota kelas tersebut
BN	:	Data diprediksi bukan anggota kelas tersebut dan memang bukan anggota kelas tersebut
SP	:	Data diprediksi anggota kelas tersebut tetapi bukan anggota kelas tersebut
SN	:	Data diprediksi bukan anggota kelas tersebut tetapi merupakan anggota kelas tersebut
Akurasi	:	% prediksi yang benar
Presisi	:	% prediksi positif yang benar
Sensitifitas	:	% baris data positif yang diprediksi sebagai positif
Spesifisitas	:	% baris data negatif yang diprediksi sebagai negative

4.3. Hasil Pengujian Base Classifier C4.5

Tabel 3 menunjukkan perbandingan akurasi classifier tunggal C4.5, metode hybrid C4.5 dan metode ensemble stacking C4.5 – naïve bayes.

Tabel 3. Hasil Perbandingan Akurasi Base Classifier C4.5

Training Dataset	C4.5 Classifier %	Hybrid C4.5 Classifier (%)	Stacking Classifier (%)
Breast cancer	75.52	81.46	71.33
Contact lenses	83.33	91.66	75
Diabetes	73.82	79.03	76.17
Glass	66.82	76.27	65.89
Iris plants	96	98.66	95.33
Soybean	91.5	92.97	93.27
Vote	96.32	97.7	96.32
Image seg.	95.73	96.53	95.67
Tic-Tac-Toe	85.07	88.1	85.18

Hasil penelitian menunjukkan tingkat akurasi stacking classfier dibandingkan dengan hybrid classifier C4.5 hanya dapat melampaui pada 2 dataset yaitu Soybean dan Tic-Tac-Toe.

4.4. Hasil Pengujian Base Classifier Naïve Bayes

Tabel 4 menunjukkan perbandingan akurasi classifier tunggal naïve bayes, metode hybrid naïve bayes dan metode ensemble stacking C4.5 – naïve bayes.

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

Tabel 4. Hasil Perbandingan Akurasi Base Classifier Naïve Bayes

Training Dataset	Naive Bayes Classifier (%)	Hybrid Naive Bayes Classifier (%)	Stacking C4.5 Classifier (%)
Breast cancer	71.67	75.87	71.33
Contact lenses	70.83	87.5	75
Diabetes	76.3	85.41	76.17
Glass	48.59	52.33	65.89
Iris plants	96	98	95.33
Soybean	92.83	94.15	93.27
Vote	90.11	94.48	96.32
Image seg.	81.06	85.19	95.67
Tic-Tac-Toe	69.62	78.91	85.18

Sementara pada perbandingan dengan metode hybrid menggunakan base classifier naïve bayes, hasil penelitian menunjukkan akurasi stacking C4.5 dan naïve bayes mlebih baik dari metode hybrid pada 4 dataset dari total 9 dataset yang digunakan.

V. KESIMPULAN

Pada penelitian ini diterapkan ensemble stacking C4.5 dan naïve bayes untuk klasifikasi multikelas. Akurasi ensemble stacking dibandingkan dengan classifier tunggal dan hybrid classifier menggunakan base classifier C4.5 dan naïve bayes. Kriteria pengukuran yang digunakan yaitu akurasi, presisi, sensitifitas dan spesifisitas. Hasil perbandingan dengan base classifier C4.5 menunjukkan ensemble stacking C4.5 dan naïve bayes hanya mampu melampui pada 2 dataset sedangkan perbandingan dengan base classifier naïve bayes menunjukkan hasil yang sedikit lebih baik pada 4 dataset.

Pada penelitian lebih lanjut akan diterapkan metode ensemble lainnya untuk klasifikasi multikelas.

REFERENSI

[1] J. Abellán and A. R. Masegosa, "Bagging schemes on the presence of class noise in classification," *Expert Syst. Appl.*, vol. 39, no. 8, pp. 6827–6837, 2012.

[2] H. Kang, S. J. Yoo, and D. Han, "Senti-lexicon and improved Na??ve Bayes algorithms for sentiment analysis of restaurant reviews," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 6000–6010, 2012.

[3] P. Duchessi and E. J. M. Lauría, "Decision tree models for profiling ski resorts' promotional and advertising strategies and the impact on sales," *Expert Syst. Appl.*, vol. 40, no. 15, pp. 5822–5829, 2013.

[4] E. Zorarpacı and S. A. Özel, "A hybrid approach of differential evolution and artificial bee colony for feature selection," *Expert Syst. Appl.*, vol. 62, pp. 91–103, 2016.

[5] R. D.Kulkarni, "Using Ensemble Methods for Improving Classification of the KDD CUP '99 Data Set," *IOSR J. Comput. Eng.*, vol. 16, no. 5, pp. 57–61, 2014.

[6] D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, and R. Strachan, "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks," *Expert Syst. Appl.*, vol. 41, no. 4 PART 2, pp. 1937–1946, 2014.

[7] M. I. Zięba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Syst. Appl.*, vol. 58, pp. 93–101, 2016.

[8] J. Quinlan, "C4. 5: programs for machine learning," *Mach. Learn.*, vol. 240, p. 302, 1993.

[9] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," *Mach. Learn.*, vol. 29, pp. 131–163, 1997.

[10] K. Polat and S. Güneş, "A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems," *Expert Syst. Appl.*, vol. 36, no. 2 PART 1, pp. 1587–1592, 2009.

[11] H. Parvin, M. Mirnabibaboli, and H. Alinejad-Rokny, "Proposing a classifier ensemble framework based on classifier selection and decision tree," *Eng. Appl. Artif. Intell.*, vol. 37, pp. 34–42, 2015.

[12] J. Abellán, "Ensembles of decision trees based on imprecise probabilities and uncertainty measures," *Inf. Fusion*, vol. 14, no. 4, pp. 423–430, 2013.

[13] Y. Chen, M. L. Wong, and H. Li, "Applying Ant Colony Optimization to configuring stacking ensembles for data mining," *Expert Syst. Appl.*, vol. 41, no. 6, pp. 2688–2702, 2014.

[14] D. H. Wolpert, "Stacked Generalization," vol. 5, no. 2, pp. 241–259, 1992.