ISBN : 978-979-587-813-1
Vol.5 No.1

CREDIT SCORING MENGGUNAKAN ALGORITMA *CLASSIFICATION AND REGRESSION TREE (CART)* DAN *ARTIFICIAL BEE COLONY*

Indra Irawan
Magister Teknik Informatika
Fakultas Ilmu Komputer Universitas Sriwijaya, Indonesia
irawanindra405@gmail.com

Dian Palupi Rini
Magister Teknik Informatika
Fakultas Ilmu Komputer
Universitas Sriwijaya, Indonesia
dian.palupi.rini@gmail.com

***Abstrak*— *Credit scoring* adalah proses penilaian kredit yang sering dilakukan oleh pihak lembaga keuangan. Melalui proses ini, ditentukan apakah calon debitur yang mengajukan kredit diklasifikasikan sebagai calon debitur yang layak untuk diberikan pinjaman atau sebaliknya. Kesalahan dalam proses *credit scoring,* pada akhirnya akan mengakibatkan kerugian dari lembaga keuangan tersebut. Kesalahan proses yang umum terjadi adalah kesalahan hasil dari prosedur credit scoring tersebut. Classification and Regression Tree merupakan salah salah satu dari sepuluh algoritma terbaik untuk digunakan di dalam data mining. kelebihan algoritma ini yang bisa mengatasi data noise. Data noise ini biasanya akan sangat sering terjadi pada data *financial.* Peneliti akan mencoba untuk menerapkan algoritma CART pada *credit scoring,* kemudian akan mencoba meningkatkan tingkat akurasi tersebut, dengan proses seleksi attribut / feature dengan menggunakan algoritma *Artificial Bee Colony* dengan menggunakan public dataset. Perbandingan akan dibuat, untuk mengetahui berapa besar kenaikan persentase akurasi menggunakan algoritma CART, sebelum dan sesudah menggunakan seleksi attribut oleh ABC.**

Abstract— Credit scoring is a credit assessment process that is often carried out by financial institutions. Through this process, it is determined whether the prospective debtor applying for credit is classified as a potential debtor to be given a loan or vice versa. Errors in the credit scoring process will ultimately result in losses from the financial institution. Common process errors are errors resulting from the credit scoring procedure. Classification and Regression Tree is one of the ten best algorithms to be used in data mining. the advantages of this algorithm can overcome data noise. This data noise usually will very often occur in financial data. Researchers will try to apply the CART algorithm to credit scoring, then try to improve the level of accuracy, by the attribute / feature selection process using the Artificial Bee Colony algorithm using a public dataset. Comparisons will be made, to find out how much the percentage increase in accuracy using the CART algorithm, before and after using the attribute selection by ABC

Kata Kunci*—*Credit Scoring, Klasifikasi, CART, ABC

I. PENDAHULUAN

Penilaian kredit adalah prosedur yang ada di setiap lembaga keuangan. Suatu cara untuk memprediksi apakah debitur memenuhi syarat untuk diberikan pinjaman atau tidak dan telah menjadi perhatian utama dalam langkah-langkah keseluruhan dari proses pinjaman. Hampir semua bank dan lembaga keuangan lainnya memiliki metode penilaian kredit mereka sendiri [1]. Saat ini, pendekatan data mining telah diterima sebagai salah satu metode yang terkenal. Tentu saja, akurasi juga merupakan masalah utama dalam pendekatan ini.

Penelitian ini mengusulkan metode algoritma CART dan Artificial Bee Colony . Indikator kinerja yang digunakan dalam penelitian ini adalah akurasi klasifikasi, tingkat kesalahan, sensitivitas, spesifisitas, dan presisi. Perbandingan akan dibuat, untuk mengetahui berapa besar kenaikan persentase akurasi menggunakan algoritma CART, sebelum dan sesudah menggunakan seleksi attribut oleh ABC.

II. DASAR TEORI

A. Algoritma Classification and Regression Tree (CART))

Algoritma Classification and Regression Tree (CART) adalah algoritma yang dikategorikan sebagai metode statistik non parametrik. Algoritma ini dikembangkan oleh Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone, yang merupakan perkembangan dan evolusi dari beberapa bidang ilmu, yaitu artificial intelligence, machine learning, non-parametric statistic, dan data mining [2].

Beberapa mekanisme yang termasuk di dalam algoritma CART adalah automatic class balancing, automatic missing value handling, cost sensitive learning, dynamic feature contruction dan probability tree estimation. Klasifikasi dan pohon regresi (CART) digunakan sebagai alat klasifikasi, di mana tujuannya adalah untuk mengklasifikasikan suatu objek menjadi dua atau lebih populasi. Seperti namanya, CART adalah satu prosedur yang dapat digunakan untuk menganalisis data kategorikal atau kontinu .Metodologi yang diuraikan dalam Breiman et al. [4] dapat diringkas menjadi tiga tahap. Pertama tahap melibatkan menumbuhkan pohon menggunakan teknik partisi rekursif untuk memilih variabel dan membagi poin menggunakan kriteria pemisahan. Metodologi CART terdiri dari 4 langkah yaitu :tree Building,Stopping Tree Building process,Tree Prunning,Optimal Tree Selection. Secara umum tree building ini, dikerjakan dua tahapan, yaitu pemilihan (classifier) penentuan simpul dan penandaan label kelas (class assignment). Pemilihan (classifier) dimulai dengan memilih metode splitter yang terbaik, yaitu dengan mengukur rata-rata dari indeks impurity dari dua child node (Lewis, 2000).

Untuk pemilihan node binary, dilakukan pengukuran gini (Gini measure of impurity) dari node t dengan persamaan:

$$G(t) = 1 - p(t)2 - (1 - p(t))2 \quad (1)$$

Dimana p(t) adalah probabilitas weight relative class 1 di node, dan improvement (gain) di generate dengan membagi node menjadi node kiri dan kanan (left and right) , sehingga didapatkan persamaan :

ISBN : 978-979-587-813-1
Vol.5 No.1

ISBN : 978-979-587-813-1
Vol.5 No.1

$$I(P) = G(P) - qG(L) - (1-q)G(R) \quad (2)$$

Dimana q probabilitas weight dari node menuju ke arah kiri. Penemu algoritma CART juga memperkenalkan teknik twoing rules, yang didasari perbandingan langsung dari attribute target di dua child node dengan persamaan :

$$I(\text{Split}) = [.25\,(q(1-q))u\,\Sigma|pL(k) - pR(k)|] \quad (3)$$

Dimana k indeks adalah target class, sedangkan pL() and pR() adalah probabilitas distibusi dari target di kiri dan kanan , dan u adalah nilai pengukuran dari pembagian child node yang tidak seimbang [2]. Tahapan terakhir berikutnya adalah penandaan class label (class label assignment). Pada tahapan ini semua node, termasuk root node dilakukan proses pelabelan class. Ini dikarenakan semua node mempunyai kesempatan yang sama sebagai node terminal[3].

B. *Artificial Bee Colony (ABC)*

ABC Algorithm adalah pendekatan metaheuristic yang diusulkan oleh Karaboga dan Basturk. Pendekatan ini terinspirasi dari perilaku cerdas kawanan lebah madu mencari makanan. Pada model ABC ini memiliki tiga kelompok lebah, yaitu Employed Bees, Onlooker Bees danScout bees[5]. Employed bee yang berhubungan dengan sumber makanan tertentu, onlooker bee menyaksikan tarian lebah yang digunakan dalam sarang untuk memilih sumber makanan, dan scout bee mencari sumber makanan secara acak. Onlooker dan scout bee merupakan unemployed bee. Awalnya scout bee menemukan posisi semua sumber makanan, setelah itu tugas dari employed bee dimulai. Sebuah employed bee buatan secara probabilitas memperoleh beberapa modifikasi pada posisi dalam memori untuk menargetkan sumber makanan baru dan menemukan jumlah nektar atau nilai fitness dari sumber baru [5]. Kemudian, scout bee mengevaluasi informasi yang diambil dari semua employed bee buatan dan memilih sumber makanan akhir dengan nilai probabilitas tertinggi terkait dengan jumlah nektar tersebut. Jika nilai fitness yang baru lebih tinggi dari yang sebelumnya, lebah itu akan melupakan yang lama dan menghafal posisi baru. Hal ini disebut sebagai greedy selection. Kemudian employed bee yang sumber makanan telah habis menjadi scout bee untuk mencari sumber makanan lebih lanjut sekali lagi.

C. *Evaluasi Model*

Evaluasi model adalah pengukuran performa model klasifikasi yang sudah dibentuk oleh classifier tertentu. Tujuannya adalah untuk mengetahui seberapa akurat, seberapa handal model klasifikasi yang terbentuk. Beberapa metrik / kriteria yang sering digunakan untuk mengukur performa dari model yang dibuat adalah accuracy / recognition rate, sensitivity/recall, specificity, precision yang dirujukan pada literatur Data Mining: Concepts and Techniques [6] .

Tabel 1. Kriteria Pengukuran Evaluasi

Metriks	Nilai
Accucary/Akurasi/Rasio Pengenalan	$\frac{TP+TN}{P+N}$
Error rate /Rasio Error/ Rasio Kesalahan Klasifikasi	$\frac{FP+FN}{P+N}$
Sensitivity / Rasio True Positive / Recall	$\frac{TP}{P}$
Specificity / Rasio True Negative	$\frac{TN}{N}$
Precision	$\frac{TP}{TP+FP}$

Tabel 2. Convolution Matriks

		Class Prediksi		
		Ya	*Tidak*	Total
Class Aktual	*Ya*	TP	FN	P
	Tidak	FP	TN	N
	Total	P'	N'	P+N

Salah satu indikator lainnya untuk menguji performa model data mining adalalah kurva ROC (Receiver Operating Characteristic). Kurva ini biasanya digunakan untuk menghitung luas area AUC (Area Under Curve). Aturan klasifikasi nilai AUC adalah sebagai berikut [4]:

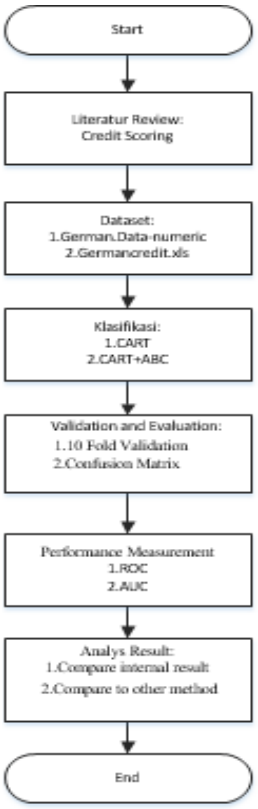
- 0,90 – 1,00 = Paling baik
- 0,80 – 0,90 = Baik
- 0,70 – 0,80 = Cukup
- 0,60 – 0,70 = Rendah
- 0,50 – 0,60 = Gagal

III. METODE PENELITIAN

Gambar 1 menunjukkan diagram alur desain penelitian yang diusulkan. Diagram alur berikut terdiri dari urutan langkah dan metode untuk melakukan penelitian. Diagram menjelaskan proses melakukan penelitian eksperimental secara lebih rinci. Peneliti akan mengikuti langkah-langkah berikut saat melakukan penelitian untuk memastikan integritas seluruh proses penelitian.

ISBN : 978-979-587-813-1
Vol.5 No.1

Gambar 1.Diagram Penelitian



Tahapan ini dilakukan dengan melakukan review dari berbagai paper yang berhubungan dengan penelitian ini. Setelah ditemukan masalah yang belum terselesaikan dari penelitian sebelumnya, maka dilakukan literature review untuk menemukan pemecahan masalah yang akan diteliti.

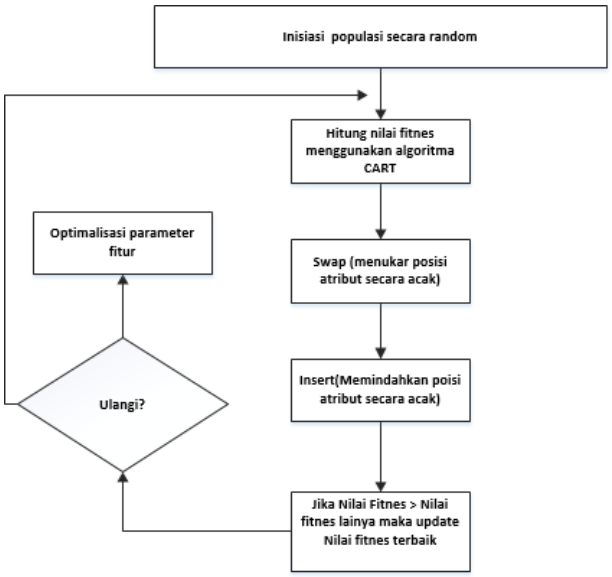
Untuk penelitian ini, dataset yang digunakan adalah german.data-numeric dan germancredit.xls. Dataset public "german.data-numeric" tersebut bisa di unduh secara online pada alamat website http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data). Sedangkan dataset public "germancredit.xls" bisa diunduh secara online pada alamat website https://ocw.mit.edu/courses/sloan-school-of management/15-062-data-mining-spring-2003assignments/GermanCredit.xls. File "german.data" terdiri dari 20 attribut prediksi, dan 1 attribut target (21 attribut)[7]. File "german.data-numeric" terdiri dari 24 attribut prediksi, dan 1 attribut target (25 attribut). Sedangkan file "germancredit.xls" terdiri dari 31 attribut dan 1 attribut target (32 attribut)[8].

Penelitian dilanjutkan dengan melakukan proses klasifikasi dengan algoritma CART dan metode yang diusulkan (CART + ABC). Prosedur eksperimental akan dilakukan dalam fase ini. Kemudian dilakukan proses validasi 10 kali lipat dan convolution matriks untuk melatih model penilaian kredit . Beberapa metrik digunakan untuk mengukur kinerja classifier. Metrik untuk mengevaluasi kinerja pengklasifikasi adalah akurasi, tingkat kesalahan, sensitivitas, spesifisitas, dan presisi. Kinerja keseluruhan ditunjukkan dalam Kurva Karakteristik Pengoperasian Penerima (ROC) dan area di bawah kurva (AUC) dari ROC [9]. Terakhir, hasil eksperimen dianalisis dan dibandingkan dengan metode data mining serupa lainnya.

Tabel 3. Dataset

Dataset	No. attribut	No. Instances
German.data-numeric	25	1000
Germancredit.xls	32	1000

Gambar 2.Metode yang diusulkan



Gambar 2 menjelaskan metode optimisasi yang diusulkan menggunakan algoritma Articial Bee Colony sebagai metode pemilihan fitur terbaik untuk menghasilkan akurasi tertinggi. Setiap atribut mewakili posisi dalam ruang biner dan posisi atribut dapat mengambil nilai biner 0 atau 1. Penelitian eksperimental ini menggunakan akurasi klasifikasi CART sebagai fungsi fitnes.Kemudian menukar posisi secara acak (swap) selanjutnya memindahkan posisi atribut secara acak(insert) ,langkah selanjutnya mengevaluasi nilai fitness terbaik yang kemudian dilakukan iterasi secara berulang sampai mendapat hasil nilai fitnes dengan akurasi terbaik[10].

IV. KESIMPULAN

Sebagai inteligen bisnis dalam pengambilan keputusan dalam hal ini credit scoring. Melalui proses ini, akan ditentukan apakah calon debitur yang mengajukan kredit diklasifikasikan sebagai calon debitur yang layak untuk diberikan pinjaman atau sebaliknya. Kesalahan dalam proses *credit scoring*, pada akhirnya akan mengakibatkan kerugian dari lembaga keuangan tersebut.Hasil penelitian ini diharapkan bisa menjadi solusi untuk meningkatkan pengambilan keputusan dalam penilaian kredit lembaga keuangan.

REFERENCES

[1] X.-L. Li, "An Overview of Personal Credit Scoring: Techniques and Future Work," Int. J. Intell. Sci., vol. 02, no. 24, pp. 182–190, 2012.

[2] S. Hussain, N. A. Dahan, F. M. Ba-alwi, and N. Ribata, "Educational Data Mining and Analysis of Students ' Academic Performance Using WEKA," Int. J. Electr. Comput. Eng., vol. 9, no. 2, pp. 447–459, 2018.

ISBN : 978-979-587-813-1
Vol.5 No.1

[3] S. M. Sadatrasoul, M. Gholamian, M. Siami, and Z. Hajimohammadi, "Credit scoring in banks and financial institutions via data mining techniques: A literature review," J. AI Data MiningJournal AI Data Min., vol. 1, no. 2, pp. 119–129, 2013.

[4] T. S. Lee, C. C. Chiu, Y. C. Chou, and C. J. Lu, "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines," Comput. Stat. Data Anal., vol. 50, no. 4, pp. 1113–1130, 2006.

[5] W. Chen, C. Ma, and L. Ma, "Mining the customer credit using hybrid support vector machine technique," Expert Syst. Appl., vol. 36, no. 4, pp. 7611–7616, 2009.

[6] J. Chen, "A Method of Improving Credit Evaluation with Support Vector Machines," 2015 11th Int. Conf. Nat. Comput., pp. 615–619, 2015.

[7] [21] M. Lichman, "{UCI} Machine Learning Repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml Expert Syst. Appl., vol. 73, pp. 1–10, 2017.

[8] Y. Ping and L. Yongheng, "Neighborhood rough set and SVM based hybrid credit scoring classifier," Expert Syst. Appl., vol. 38, no. 9, pp. 11300–11304, 2011.

[9] F. Gorunescu, "Classification Performance Evaluation," in Data Mining. Concepts, Models and Techniques, vol. 12, 2011, pp. 319–330.

[10] A. Bandhu and M. Kumar, "Computational time reduction for credit scoring : An integrated approach based on support vector machine and stratified sampling method," Expert Syst. Appl., vol. 39, no. 8, pp. 6774–6781, 2012.

ISBN : 978-979-587-813-1
Vol.5 No.1