PARTICLE SWARM OPTIMIZATION MENINGKATKAN AKURASI NAÏVE BAYES CLASSIFIER

Suamanda Ika Novichasari

Universitas Dian Nuswantoro
Email : vichareal0311@gmail.com

ABSTRAK

Salah satu teknik klasifikasi data mining adalah *Naïve Bayes Classifier* (NBC) , namun hasil akurasinya masih kurang dibanding algoritma C4.5 dan *neural network*. NBC unggul jika diterapkan pada data ukuran besar, namun lemah pada seleksi atribut. Artikel ini berisi tentang penggunaan algoritma *Particle Swarm Optimizatin* (PSO) untuk membobot atribut guna meningkatkan nilai akurasi NBC. Penelitian ini menggunakan *data set* publik *German Credit Data*. Proses validasi menggunakan *tenfold-cross validation*, sedangkan pengujian modelnya menggunakan *confusion matrix* dan kurva ROC. Hasilnya menunjukan akurasi NBC meningkat dari 73,70% menjadi 78,00% setelah dikombinasikan dengan PSO.

Kata kunci : Kelayakan kredit, data mining, teknik klasifikasi data mining, NBC, NBC-PSO.

1. PENDAHULUAN

Data mining adalah suatu proses yang bertujuan untuk menemukan pola secara otomatis atau semi otomatis dari data yang sudah ada di dalam basis data yang dimanfaatkan untuk menyelesaikan suatu masalah [1]. *Data mining* memiliki beberapa teknik, diantaranya klasifikasi dan *clustering*. Teknik klasifikasi adalah teknik pembelajaran yang digunakan untuk memprediksi nilai dari atribut kategori target [2]. Klasifikasi bertujuan untuk membagi objek yang ditugaskan hanya ke salah satu nomor kategori yang disebut kelas [3]. *Clustering* mengelompokkan objek atau data berdasarkan kemiripan antar data, sehingga anggota dalam satu kelompok memiliki banyak kemiripan dibandingkan dengan kelompok lain [4]. Untuk menyelesaikan masalah analisa resiko kredit data akan diklasifikasikan menjadi dua kelas, yaitu kredit baik dan kredit buruk. Sehingga tepat menggunakan teknik klasifikasi *data mining*. Metode yang paling populer digunakan untuk teknik klasifikasi adalah *Decision Trees*, *Naïve Bayes Classifiers* (NBC), *Statistical analysis*, dan lain lain [4].

Dari hasil penelitian Henny Leidiyana [5] algoritma NBC untuk kelayakan kredit hasil akurasinya masih kurang dibanding menggunakan algoritma C4.5. Dalam C4.5 seluruh atribut diseleksi untuk kemudian dibagi menjadi himpunan bagian yang lebih kecil, namun jika data berukuran besar dengan banyak atribut maka model yang terbentuk menjadi rumit dan sulit dipahami, sehingga perlu dilakukan pemangkasan (*pruning*) yang dapat mengurangi akurasi. Sedangkan NBC lebih tepat diterapkan pada data yang besar [6]. Dapat menangani data yang tidak lengkap (*missing value)* serta kuat terhadap atribut yang tidak relevan dan *noise* pada data [4]. NBC akan bekerja lebih efektif jika dikombinasikan dengan beberapa prosedur pemilihan atribut [1].

2. NAÏVE BAYES

Disebut juga dengan Bayesian *Classification* adalah pengklasifikasian statistik yang didasarkan pada teorema bayes yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. Bayesian *Classification* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database yang besar [7].

Bentuk umum teorema bayes sebagai berikut :

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \dots\dots\dots 2.1$$

Dimana :

- X = data dengan kelas yang belum diketahui
- H = Hipotesa data X merupakan suatu kelas spesifik
- P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X (*posterior probability*)
- P(H) = probabilitas hipotesis H (*prior probability*)

Peluang bersyarat atribut kategorikal dinyatakan dalam bentuk [4]:

$$P(A_i|C_j) = \frac{|A_{ij}|}{N_{C_j}} \dots\dots\dots\dots 2.3$$

Dimana $|A_{ij}|$ adalah jumlah contoh *training* dari kelas A_i yang menerima nilai C_j. Jika hasilnya adalah nol, maka menggunakan pendekatan yang dinyatakan dalam bentuk [8]:

$$P(A_i|C_j) = \frac{n_c + n_{equiv}\, p}{n + n_{equiv}} \dots\dots\dots\dots 2.4$$

Dimana *n* adalah total dari jumlah *record* dari kelas C_j. n_c adalah jumlah contoh *training* dari kelas A_i yang menerima nilai C_j. n_{equiv} adalah nilai konstan dari ukuran sampel yang equivalen. *P* adalah peluang estimasi prior, *P=1/k* dimana k adalah jumlah kelas dalam variable target.

Peluang bersyarat atribut kontinu dinyatakan dalam bentuk [4]:

$$P(A_i|C_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}}\, exp\left[-\frac{(A_i-\mu_{ij})^2}{2\,(\sigma_{ij})^2}\right] \dots\dots\dots\dots 2.5$$

Parameter μ_{ij} dapat diestimasi berdasarkan sampel mean A_i untuk seluruh training record yang dimiliki kelas C_j. Dengan cara sama, $\sigma^2{}_{ij}$ dapat diestimasi dari sampel varian $\left(s^2\right)$ training record tersebut.

3. PARTICLE SWARM OPTIMIZATION

PSO adalah metode optimasi heuristic global yang diperkenalkan oleh Dokter Kennedy dan Eberhart pada tahun 1995 berdasarkan penelitian terhadap perilaku kawanan burung dan ikan [9].

Setiap partikel dalam PSO juga dikaitkan dengan kecepatan partikel terbang melalui ruang pencarian dengan kecepatan yang dinamis disesuaikan untuk perilaku historis mereka. Oleh karena itu, partikel memiliki kecenderungan untuk terbang menuju daerah pencarian yang lebih baik dan lebih baik selama proses pencarian [10].

Rumus untuk menghitung perpindahan posisi dan kecepatan partikel yaitu [11]:

$$V_i(t) = V_i(t-1) + c_1 r_1\left[X_{pbest_i} - X_i(t)\right] + c_2 r_2[X_{Gbest} - X_i(t)] \dots 2.6$$

$$X_i(t) = X_i(t-1) + V_i(t) \dots\dots 2.7$$

Dimana :

- Vi (t) = kecepatan partikel *i* saat iterasi *t*
- X_i (t) = posisi partikel *i* saat iterasi *t*
- c_1 dan c_2 = *learning rates* untuk kemampuan individu (cognitive) dan pengaruh sosial (group)
- r_1 dan r_2 = bilangan random yang berdistribusi uniformal dalam interval 0 dan 1
- XPbesti = posisi terbaik partikel i
- XGbest = posisi terbaik global

4. NAÏVE BAYES BERBASIS PARTICLE SWARM OPTIMIZATION

PSO diterapkan pada pembobotan atribut seperti algoritma dibawah ini :

- Identifikasi populasi sample
- Hitung $P(C_i)$ pada setiap kelas
- Inisialisasi posisi setiap patikle
- For each atribut do
  - Evaluasi nilai fungsi tujuan
  - Cari Pbest dan Gbest
  - Update kecepatan dan posisi particle
  - Gbest = bobot atribut ke-j
- hitung P(X|Ci), i=1,2 untuk setiap kelas atau atribut
- $P(X|C_1) > P(X|C_2)$?

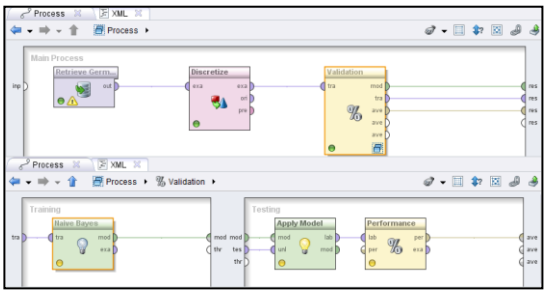
Data dari atribut numerik diubah menjadi nominal, kemudian identifikasi populasi sampel dari *data set.* Hitung $P(C_i)$ untuk setiap kelas, dalam kasus *data set* pada penelitian ini terdiri dari 2 kelas yaitu kredit baik yang dinyatakan dengan "1" dan kredit buruk yang dinyatakan dengan "2".

Inisialisasi posisi setiap partikel atribut ke-j merupakan awal dari tahap pembobotan atribut dengan PSO. Langkah selanjutnya adalah evaluasi nilai fungsi tujuan dari setiap partikel untuk mendapatkan posisi terbaik (Pbest) dan posisi global terbaik (Gbest), kemudian update kecepatan dan posisi partikel. Ulangi langkah evaluasi nilai fungsi tujuan sampai mencapai konvergen, kemudian Gbest = bobot atribut ke-j. Cek apakah nilai j sudah maksimal, jika belum ulangi langkah-langkah dari inisialisasi posisi setiap partikel atribut ke-j sampai menemukan bobot atribut ke-j. Ulangi langkah tersebut sampai nilai j sudah maksimal atau semua atribut sudah terbobot.

Kemudian hitung P(X|Ci), i=1,2 untuk setiap kelas atau atribut. Setelah itu bandingkan, jika $P(X|C_1) > P(X|C_2)$ maka kesimpulannya adalah C_1 atau dalah kasus pada penelitian ini bearti kredit baik. Jika $P(X|C_1) < P(X|C_2)$ maka kesimpulannya C_2 atau kredit buruk.

5. EKPERIMEN

Data yang digunakan pada penelitian ini berasal dari University of California, Irvine (UCI) Machine Learning dengan judul German Credit data. Data ini berjumlah 1000 *record* dan terdiri dari 20 atribut, dengan 7 atribut bertipe numerik dan 13 bertipe kategorikal [12].

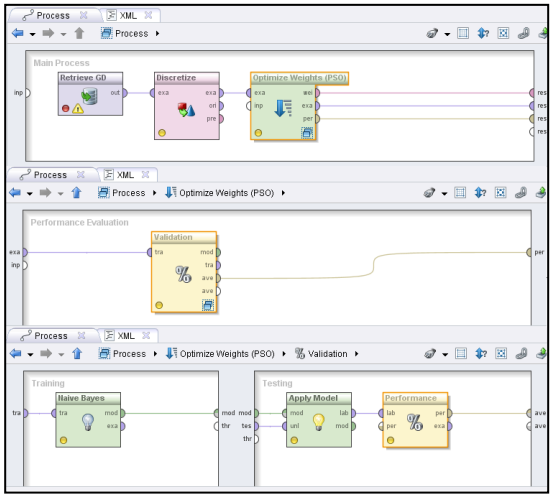


Gambar 1. Desain model NBC

Hasil dari model di atas menghasilkan nilai akurasi confusion

matrix sebesar 73,70% dan akurasi AUC 0,774 dalam selang waktu 1 detik.

Untuk NBC-PSO, pertama kali dilakukan uji coba dengan memberi nilai pada parameter *population size* antara 10-600 dan *maximum number of generation* 100 bernilai konstan. *Population size* adalah jumlah individual pada tiap generasi, sedangkan *maximum number of generation* adalah jumlah generasi maksimum untuk menghentikan jalannya algoritma. Terpilih nilai *population size* terbaik adalah 350 dengan hasil akurasi 77,80 % dan AUC 0,771.



Gambar 1. Desain model NBC-PSO

Selanjutnya dilakukan percobaan dengan population size bernilai tetap 350 dan maximum number of generation bernilai 100-1500. Akurasi tertinggi dan waktu eksekusi terendah terjadi pada saat *maximum number of generation* bernilai 500 dengan nilai akurasi sebesar 78,00%, AUC 0,778 dalam waktu 2 jam 6 menit 49 detik.

6. HASIL

Berdasarkan hasil percobaan, diperoleh akurasi NBC-PSO paling tinggi terjadi pada saat *population size* bernilai 350 dan *maximum number of generation* bernilai 500. Akurasi NBC-PSO 78,00%, dan AUC 0,778 sedangkan akurasi NBC hanya 73,70% dan AUC 78,00%.

Tabel 1. Komparasi akurasi NBC dan NBC-PSO

Perbandingan	NBC	NBC-PSO
Akurasi confusion matrix (%)	73,70	78,00
Akurasi AUC	0,774	0,778
Waktu eksekusi	1 s	2 h. 6 m. 49 s

Tabel 1. Komparasi akurasi NBC dan NBC-PSO

Atribut	Bobot
status of existing checking account	0.519
duration in mounth	1
credit history	1
Purpose	0
credit amount	0
savings account	1
present employment since	1
instalment of disposable income	1
personal status n sex	0
other debtors/guarantors	1
Present residence since	0
Property	1
Age	1
Other installment plans	0
Housing	0
existing credits at this bank	1
Job	0
number of people being liable to provide maintenance for	1
Telephone	1
foreign work	1

Hasil pembobotan atribut yaitu 7 atribut mempunyai bobot 0, 12 atribut mempunyai bobot 1 dan 1 atribut mempunyai bobot 0,519. Sehingga atribut yang berbobot 0 dapat dihilangkan karena tidak mempunyai pengaruh pada akurasi kelayakan kredit bank.

7. KESIMPULAN

Hasil percobaan membuktikan bahwa PSO yang diterapkan pada pembobotan atribut meningkatkan akurasi NBC. Akurasi meningkat 4,30% dan AUC meningkat 0,004.

Dengan demikian terbukti bahwa PSO yang diterapkan pada pembobotan atribut NBC meningkatkan nilai akurasi. Hal ini menjadikan NBC-PSO memberikan pemecahan untuk permasalahan kelayakan kredit bank lebih akurat.

8. DAFTAR PUSTAKA

[1] Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools And Techniques.* Burlington, Usa: Morgan Kaufmann Publishers.

[2] Vercellis, Carlo (2009). *Business Intelligent: Data Mining and Optimization for Decision Making*. Southern Gate, Chichester, West Sussex: John Willey & Sons, Ltd.

[3] Bramer, Max. (2007). Principles of Data Mining. London: Springer.

[4] Gorunescu, F. (2011). Data Mining Concepts,Models And Techniques. Verlag Berlin Heidelberg: Springer.

[5] Leidiyana, H (2012). Komparasi Algoritma Klasifikasi Data Mining Dalam Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor. Tesis Magister Ilmu Komputer. Sekolah Tinggi Managemen Informatika dan Komputer Nusa Mandiri.

[6] Wu, Xindong and Kumar, Vipin. (2009). The Top Ten Algorithms in Data Mining. Boca Raton: CRC Press.

[7] Kusrini,&Luthfi, E. T. (2009).Algoritma Data Mining. Yogyakarta: Andi Publishing.

[8] Larose, D. T. (2006).Data Mining Methods and Models. New Jersey: John Willey & Sond, Inc.

[9] J. Kennedy and R. C. Eberhart. Particle swarm optimization. In Proceedings of the 1995 IEEE International Conference on Neural Networks. IEEE Service Center, Piscataway, 1995.

[10] Abraham, A., Grosan, C., & Ramos, V. (2006). Swarm Intelligence In Data Mining. Verlag Berlin Heidelberg: Springer.

[11] Lin, J dan Yu, J (2009). Weighted Naïve Bayes classification algorithm based on particle swarm optimization. Yunnan University of Finance and Economics Yunnan Kunming, China.

[12] http://archive.ics.uci.edu/ml/datasets /Statlog+(German+Credit+Data), di akses pada tanggal 26 Maret 2013.