PENCARIAN MODEL TERBAIK ANTARA ALGORITMA C4.5 DAN C4.5 BERBASIS PARTICLE SWARM OPTIMIZATION UNTUK PREDIKSI PROMOSI DEPOSITO

Syaeful Mujab

NIM : A11.2009.04895

Program Studi Teknik Informatika

Fakultas Ilmu Komputer

Universitas Dian Nuswantoro, Jl. Nakula 5-11, Semarang

Email :syaeful.mujab1@gmail.com

ABSTRAK

Keberhasilan promosi atau pemasaran deposito pada sebuah bank sangat berperan dalam meningkatkan dan menjaga kelangsungan hidup sebuah bank. Oleh karena itu sangat penting untuk mengetahui kelompok atau nasabah yang berpotensi melakukan deposito atau tidak. Dari kondisi tersebut teknik data mining yang tepat digunakan adalah klasifikasi. Salah satu teknik klasifikasi data mining adalah C4.5 , algoritma C4.5 mempunyai keunggulan dalam kecepatan membaca dan membentuk model sehingga mudah dipahami, namun mempunyai kelemahan dalam pembacaan data yang berjumlah besar. Laporan ini menggunakan algoritma pembobotan Particle Swarm Optimizatin (PSO) dengan seleksi atribut guna meningkatkan akurasi C4.5. Desain penelitian menggunakan model proses CRISP-DM karena penyelesaian masalah dalam penelitian ini mengarah pada masalah strategi bisnis. Data yang digunakan dalam penelitian ini adalah public dataset bank portugis. Proses validasi menggunakan tenfold-cross validation, pengujian menggunakan model confusion matrix dan kurva ROC. Hasil akurasi C4.5 setelah dikombinasi dengan PSO terbukti meningkat dari 88.83% menjadi 89,26%.

***Kata kunci : deposito, pemasaran, data mining, teknik klasifikasi data mining, C4.5, C4.5-PSO*.**

1. Pendahuluan

1.1. Latar Belakang Masalah.

Di era yang kompetitif seperti saat ini promosi dan pemasaran produk atau jasa dalam suatu perusahaan sangatlah penting, pemasaran digunakan untuk memperkenalkan atau menjual produk dari perusahaan kepada konsumen. Promosi atau pemasaran adalah upaya untuk memberitahukan atau menawarkan produk atau jasa pada dengan tujuan menarik calon konsumen untuk membeli atau mengkonsumsinya. Dengan adanya promosi produsen atau distributor mengharap kenaikan angka penjualan.

Bagi perusahaan ada dua cara pendekatan utama yang digunakan untuk mempromosikan produk atau jasa: melalui kampanye massal, target umum(bersifat acak) atau pemasaran terarah, memilih target lebih spesifik berdasarkan kriteria yang telah dibuat (Ling dan Li 1998). Namun saat ini, tanggapan positif terhadap kampanye massa biasanya sangat rendah, kurang dari 1%, menurut studi yang sama. Atau, fokus pemasaran diarahkan pada target yang perkirakan akan lebih spesifik dengan produk / layanan tertentu, membuat kampanye semacam ini lebih

menarik karena lebih efisien (Ou et al. 2003). Namun demikian, pemasaran yang diarahkan juga memiliki beberapa kelemahan, misalnya dapat memicu sikap negatif terhadap bank karena masalah privasi seseorang (Page and Luding 2003).

Data Mining(DM) adalah teknologi BI yang menggunakan model data-driven untuk mengekstrak pengetahuan yang berguna (misalnya pola) dari data yang kompleks dan luas (Witten dan Frank, 2005). The CRoss-Industry Standard Process for Data Mining (CRISP-DM) adalah metodologi populer untuk meningkatkan keberhasilan proyek DM (Chapman et al., 2000). Metodologi ini mendefinisikan urutan prosesnya menjadi enam fase, yang memungkinkan pelaksanaan pembangunan model DM untuk digunakan dalam lingkungan yang nyata, membantu untuk mendukung keputusan bisnis. Beberapa metode pada Data Mining yang terkait dalam penelitian strategi pemasaran langsung antara lain NB(Naïve Bayes) (Zhang, 2004), DT(Decision Trees) (Aptéa and Weiss, 1997) SVM(and Support Vector Machines) (Cortes and Vapnik, 1995).

2. Landasan Teori

2.1. Pengertian Data Mining

Data mining adalah analisis data(sering besar) pengamatan dataset untuk menemukan hubungan tidak terduga dan untuk meringkas data dengan cara baru yang baik dimengerti dan berguna untuk pemilik data (Hand et al.). Data mining adalah bidang interdisipliner yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, basis data, dan visualisasi untuk mengatasi masalah ekstraksi informasi dari basis data yang besar(EvangelosSimoudis in Cabena et al).

2.2. Algoritma C4.5

Algoritma C4.5 di temukan oleh Quinlan merupakan turunan dari algoritma sebelumnya yaitu ID3 yang sama-sama menghasilkan pohon keputusan. Sama seperti dengan CART, algoritma C4.5 rekursif mengunjungi setiap node keputusan, dan memilih cabang yang paling optimal, sampai tidak ada lagi cabang yang mungkin dikunjungi.

Algoritma C4.5 dalam sebuah membuat keputusan mempunyai tahapan sebagai berikut: (Gorunescu, 2011) yaitu:

1. *Siapkan data training ,bisa dari histori sebelumnya dan sudah dikelompokkan menurut kelasnya.*
2. Menentukan akar dari pohon dengan menghitung nilai *gain* yang tertinggi dari masing-masing atribut atau berdasarkan nilai *index entropy* terendah. Sebelumnya dihitung terlebih dahulu nilai *index entropy*, denganrumus:

$$entropy(i) = -\sum_{j=1}^{m} f(i,j).\log 2\, f[(i,j)]$$

Keterangan:

i = himpunan kasus

m = jumlah partisi i

f(i,j) = proposi j terhadap i

3. Hitung nilai *gain* dengan rumus:

$$Entropysplit = -\sum_{i=1}^{p}.1E(i)$$

Keterangan:

p = jumlah partisi atribut

ni = proporsi ni terhadap i

n = jumlah kasus dalam n

4. Ulangi langkah ke-2 hingga semua *record* terpartisi

Proses partisi pohon keputusan akan berhenti disaat:

a. Semua tupel dalam *record* dalam simpul m mendapat kelas yang sama

b. Tidak ada atribut dalam *record* yang dipartisi lagi tidak ada *record* didalam cabang yang kosong.

2.3. Algoritma PSO

Disebut Algortima Particle Swarm Optimizatio(PSO) terinspirasi sebuah perilaku cerdas burung dan ikan dalam mencari makan. Ada teori PSO yang menyatakan bahwa proses adaptasi budaya berakar dalam tiga prinsip: mengevaluasi,membandingkan dan meniru. Dari prinsip inilah algoritma PSO dianggap sebagai algoritma yang cerdas karena mampu membandingkan sebelum mengeksekusi.

Modifikasi kecepatan dan posisi tiap partikel dapat dihitung menggunakan kecepatan saat ini dan jarak pbesti, d ke gbestd seperti ditunjukan persamaan berikut:

$$vi,d = w*vi,d + c1*R*(pbesti,d - xi,d) + c2*R*(gbestd - xi,d)$$

$$xi,d = xi,d + vi,d$$

Dimana:

Vi,d = Kecepatan partikel ke-i pada iterasi ke-i

w = Faktor bobot inersia

c1, c2 = Konstanta akselerasi (learning rate)

R = Bilangan random (0-1)

xi,d = Posisi saat ini dari partikel ke-i pada iterasi ke-i

pbesti = Posisi terbaik sebelumnya dari partikel ke-i

gbesti = Partikel terbaik diantara semua partikel dalam satu kelompok atau populasi

n = Jumlah partikel dalam kelompok d = Dimensi

Persamaan (2.3) menghitung kecepatan baru untuk tiap partikel (solusipotensial) berdasarkan pada kecepatan sebelumnya (Vi,m), lokasi partikel dimana nilai fitness terbaik telah dicapai (pbest), dan lokasi populasi global (gbest untuk versi global, lbest untuk versi local) atau local neighborhood pada algoritma versi local dimana nilai fitness terbaik telah dicapai. Persamaan (2.4) memperbaharui posisi tiap partikel pada ruang solusi. Dua bilangan acak c1 dan c2 dibangkitkan sendiri. Penggunaan berat inersia w telah memberikan performa yang meningkat pada sejumlah aplikasi. Hasil dari perhitungan partikel yaitu kecepatan partikel diantara interval [0,1] (Hu, Shi, &Eberhart, 2004).

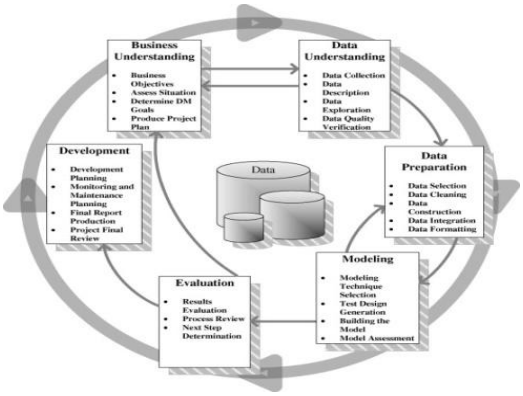
3. Metode Penelitian

3.1 Sumber Data

Pada penelitian ini data yang digunakan berasal dari public dataset University of California, Irvine (UCI) Machine Learning. Data tersebut pernah digunakan oleh S. Moro, R. Laureano and P. Cortez , yang berjudul Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.

3.2 Metode Penelitian

CRIPS-DM(Cross-Industry Standart Proses for Data Mining) dikembangkan pada tahun 1996 oleh analis dari beberapa industri. CRIPS-DM menyediakan standart proses data mining sebagai pemecahan masalah secara umum dari bisnis atau unit penelitian. CRIPS-DM memiliki siklus hidup yang terbagi dalam enam fase, yaitu:

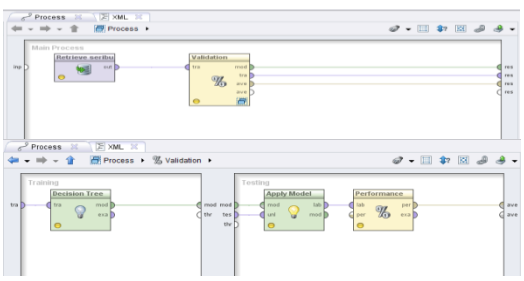


Gambar 3.1 Tahapan Proses *CRISP-DM*(Larose. 2005)

4. Pembahasan

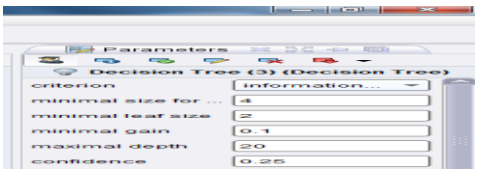
4.1. Validasi dan Evaluasi

4.1.1 Desain model C4.5



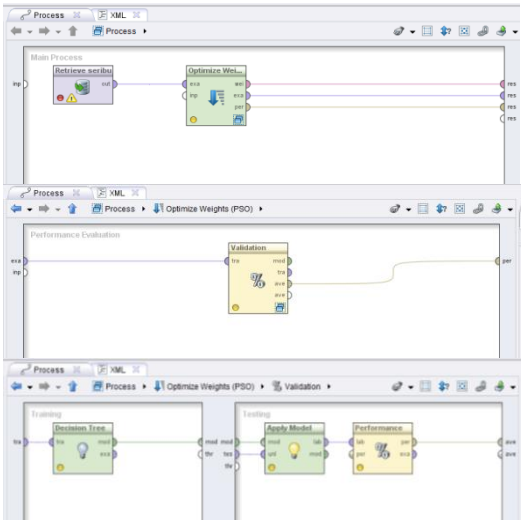
Gambar 4.1 desain model C4.5

4.1.2 Pengaturan parameter



Gambar 4.2 pengaturan parameter *decision tree* pada *rapidminer*

4.1.3 Desain model C4.5-PSO



Gambar 4.3 Desain model validasi C4.5 berbasis PSO

4.2. Hasil Pengujian dan Analisis

4.3.1 Hasil Pengujian

Tabel 4.1 perbandingan akurasi

Jumlah Data	*C4.5*		*C4.5-PSO*	
	akurasi	*AUC*	*akurasi*	*AUC*
99	*96%*	*0.500*	*98%*	*0.500*
999	*94.89%*	*0.908*	*95.70%*	*0.920*
7467	*87.53%*	*0.883*	*88.57%*	*0.912*
25000	*95.06%*	*0.925*	*95.18%*	*0.925*
45211	*88.83%*	*0.868*	*89.26%*	*0.874*

4.3.2 Analisa Hasil Pengujian

Percobaan pada penelitian ini menggunakan RapidMiner 5.3.008. Algoritma yang digunakan adalah C4.5 dan C4.5-PSO untuk pembobotan atribut. Validasinya menggunakan tenfold cross-

validation, sedangkan pengukuran performanya menggunakan confusion matrix dan kurva ROC.

Nilai dari population size dan maximum number of generation pada PSO diubah-ubah untuk meningkatkan kinerja PSO yang berdampak pada peningkatan akurasi.

Berdasarkan hasil percobaan, diperoleh akurasi C4.5-PSO 89,26%, dan AUC 0,874 sedangkan akurasi C4.5 hanya 88.83% dan AUC 0.868.

5. Penutup

5.1. Kesimpulan

Pada penelitian ini dilakukan pemodelan menggunakan algoritma C4.5 dan C4.5 yang dikombinasi dengan PSO, data yang digunakan adalah data nasabah bank yang mana pada tujuan ini untuk mengetahui nasabah mana yang nantinya berpotensi melakukan deposito. Penelitian ini difokuskan pada penerapan algoritma PSO sebagai pembobotan atribut teknik klasifikasi data mining C4.5.Validasi model menggunakan 10fold cross-validation dan evaluasi model menggunakan confusion matrix dankurva ROC.

Dari penelitian ini didapat algoritma C4.5 yang dikombinasi dengan algoritma PSO mempunyai akurasi yang lebih baik dibanding penggunaan algoritma C4.5 saja, yaitu 88.83% berbanding 89.26% akan tetapi untuk waktu eksekusi algoritma C4.5 yang dikombinasi PSO memakan waktu lebih lama yaitu 6 menit 44 detik untuk C4.5 dan menjadi 4 jam 48 menit 44 detik untuk C4.5 berbasis PSO.

Terbukti algoritma PSO yang digunakan sebagai algoritma pembobot mampu meningkatkan akurasi dari algoritma C4,5. Sehingga C4.5-PSO bisa dijadikan solusi untuk mengetahui nasabah mana yang berpotensi melakukan deposito.

5.2. Saran

1. Tools bantu berupa rapid miner yang digunakan adalah versi 5.3.008, untuk mendapat kan hasil yang lebih baik bisa menggunakan versi terbaru karena sekarang sudah ada versi 5.3.13.
2. Dataset yang digunakanp ada penelitian ini adalah public dataset yang diambil dari archive UCI yaitu data nasabah bank yang ada di Portugal, untuk penelitian selanjutnya bisa menggunakan data nasabah bank di Indonesia.
3. Penelitian ini mengkomparasikan algoritma C4.5 dan C4.5 yang dikombinasikan dengan PSO sebagai pembobotan atribut, untuk penelitian selanjutnya dapat dikembangkan dengan menggunakan algoritma klasifikasi lain seperti Suport Vector Machine (SVM), Neural Network, Nieve Bayes yang dikombinasikan dengan algoritma Adaboost, atau algoritma optimasi lain seperti Ant Colony Optimization (ACO), Genetic Algorithm (GA), PSO atau algoritma optimasi lainnya.
4. Hardware yang digunakan dalam penelitian ini sangat terbatas yaitu RAM 2 GB, prosesor dual core dan untuk selanjutnya bisa digunakan hardware yang lebihbaik yaitu menggunakan RAM 8 GB dan prosesor intel i7.

DAFTAR PUSTAKKA

[1] *[Witten, H. I., Frank, E., & Hall, M. A. (2011). Data Mining Pratical Mechine Learning Tools And Technique. Burlington: Elsevier Inc.*

[2] *Larose, D. T. (2005).Discovering Knowledge in Data. New Jersey: John Willey & Sons, Inc.*

[3] *Sousa, T., Silva, A., Neves, A. (2004).Particle Swarm based Data Mining Algorithms for classification tasks.Elsevier.*

[4] *Saleszberg S L. Book Review: C4.5:by J. Ross Quinlan.Inc., 1993.*
Programs for Machine Learning Morgan Kaufmann Publishers.

[5] *Hariyanto, S (2012). Segmentasi dan Klasifikasi Perilaku Pembayaran Pelanggan pada Perusahaan Penyedia Layanan Multimedia denganAlgoritma KMeans dan C4.5. Thesis Magister Ilmu Komputer. Sekolah Tinggi Manajemen Informatika dan Komputer Eresha.*

[6] *Irfiani, E.(2011). Penerapan Algoritma Klasifikasi C4.5 Berbasis AdaBoost Untuk Prediksi Loyalitas Pelanggan Distributor Pulsa Elektronik. Thesis Magister Ilmu Komputer.*

[7] *Lasut, D. (2012). Prediksi Loyalitas Pelanggan Pada Perusahaan Penyedia Layanan Multimedia Dengan Algoritma C4.5 Berbasis Particle Swarm Optimization. Thesis Magister Ilmu Komputer. Sekolah Tinggi Manajemen Informatika dan Komputer Eresha.*

[8] *Moro, S., Laureano, M.S., Cortez, P. (2011)Using Data Mining For Bank Direct Marketing:An Application Of The CRISP-DM Methodology. Proceedings of the European Simulation and Modelling Conference.*

[9] *C. Apte and S.M. Weiss (1997). Data Mining with Decision Trees and Decision Rules. Future Generation Computer Systems.*

[10] *Rocha, B, C., Junior, R, T, S. (2010) Identifying Bank Frauds Using CRISP-DM And Decesion Trees. International journal of computer science & information Technology.*

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

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

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

[14] *Ling,C, X. &Li, C(1998). Data Mining for Direct Marketing: Problems and Solutions.*

[15] *Vapnik, V. and Cortes, C(1995). Support-Vector Network.*

[16] *Tsai, C. F.,& Chen, M. Y(2009). Variable Selection by Association Rules for Customer Churn Prediction of Multimedia on Demand Expert Systems with Application.*

[17] *Zhang, H(2004). The Optimality of Naïve Bayes.*

[18] *http://archive.ics.uci.edu/ml/datasets/Bank+Marketing, diakses tanggal 23 September 2013*