ANALISIS DAN IMPLEMENTASI KLASIFIKASI DATA MINING MENGGUNAKAN JARINGAN SYARAF TIRUAN DAN EVOLUTION STRATEGIES

Naufar Rifqi[1], Warih Maharani[2], Shaufiah[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Data mining merupakan suatu proses ekstraksi informasi yang berguna dari sekumpulan data yang terdapat secara implisit dalam suatu basis data. Metode untuk melakukan ekstrasi informasi tersebut salah satunya adalah klasifikasi. Tujuan klasifikasi adalah untuk menganalisa data input dan membuat model yang akurat untuk setiap kelasnya berdasarkan data yang ada. Model kelas tersebut juga digunakan untuk mengklasifikasikan data tes lain. Jaringan Syaraf Tiruan (JST) merupakan salah satu algoritma yang digunakan untuk melakukan klasifikasi. Kelebihan dari JST yaitu memiliki aturan pelatihan (training rule) untuk menemukan bobot-bobot koneksi berdasarkan data latih dalam pembelajaran (learning) mengenali pola. Sehingga, JST akan dapat mengenali pola dengan akurasi tinggi jika sudah melakukan proses pembelajaran dan memiliki arsitektur yang optimal. Dalam mencari arsitektur yang optimal pada JST bukanlah hal yang mudah. Sehingga diperlukan algoritma optimasi agar dapat memperoleh arsitektur yang optimal. Pada tugas akhir ini digunakan Evolution Strategies (ES) yang merupakan algoritma optimasi berbasis evolusi. Pada tugas akhir ini digunakan dua buah data set, yaitu : data Pima Indians Diabetes dan Breast Cancer. Data-data dibagi menjadi tiga bagian menjadi data training, data validation dan data testing. Dengan gabungan metode JST dan ES dapat menghasilkan suatu sistem klasifikasi yang akurat khususnya pada saat testing dengan diperoleh fungsi fitness 81,2834 % untuk data Pima Indians Diabetes dan 98,2456 % untuk data Breast Cancer.

Kata Kunci : Data Mining, Klasifikasi, Jaringan Syaraf Tiruan, Evolution Strategies

Abstract

Data mining is a process of extracting useful information from data sets implicitly contained within the database. Method for extraction of such information is one of classification. The purpose of classification is to analyze the input data and create accurate models for each class based on existing data. Class model is also used to classify the data of other tests. Artificial Neural Network (ANN) is one of the algorithms used to perform classification. The advantages of ANN is a training rule to find the connection weights based on training data in learning (learning) to recognize patterns. Thus, the ANN will be able to recognize patterns with high accuracy if it is doing the learning process and have the optimal architecture. In searching for the optimal architecture of ANN is not easy. So that is needed for the optimization algorithm to obtain an optimal architecture. In this final use Evolution Strategies (ES) which is the evolution-based optimization algorithm. In this thesis used two data sets, namely: data Pima Indians Diabetes and Breast Cancer. The data is divided into three sections into training data, data validation and testing data. With the combination of ANN and ES method can yield an accurate classification system, especially at the time of testing with the fitness function obtained 81.2834% for Pima Indians Diabetes data and 98.2456% for Breast Cancer data.

Keywords : Data Mining, Classification, Artificial Neural Networks, Evolution Strategies

1 Pendahuluan

1.1 Latar Belakang

Berkembangnya penggunaan komputer dalam bidang manajemen data menyebabkan akumulasi data dalam jumlah sangat besar di beberapa organisasi. Apalagi dengan berkembangnya persepsi bahwa analisa terhadap data yang besar ini akan mengubah data pasif menjadi informasi yang berguna. Salah satu cara untuk melakukan hal itu adalah dengan menggunakan metode Data Mining atau *Knowledge Discovery in Databases*. Di dalam konsep Data Mining terdapat berbagai cara dan metode untuk mengekstrak informasi dari data yang besar.

Klasifikasi adalah salah satu metode untuk melakukan ekstraksi informasi. Data input atau yang biasa disebut dengan *training set*, terdiri dari banyak *record*, yang tiap record-nya mempunyai beberapa atribut. Setiap record ini juga mempunyai sebuah label kelas. Tujuan dari klasifikasi ini adalah untuk menganalisa data *input* dan membuat model yang akurat untuk setiap kelasnya berdasarkan data yang ada. Model kelas tersebut juga digunakan untuk mengklasifikasikan data tes lain/data tes baru untuk ditentukan label kelasnya.

Terdapat banyak algoritma yang dapat digunakan untuk melakukan klasifikasi data mining, salah satunya dengan Jaringan Syaraf Tiruan (JST). JST merupakan suatu arsitektur jaringan untuk memodelkan cara kerja sistem syaraf manusia (otak) dalam melaksanakan tugas tertentu. Pemodelan ini didasari oleh kemampuan otak manusia dalam mengorganisasi sel-sel penyusunnya atau *neuron*, sehingga memiliki kemampuan untuk melaksanakan tugas-tugas tertentu khususnya pengenalan pola dengan efektivitas jaringan sangat tinggi [9]. Dalam mencari arsitektur yang optimal bukanlah hal yang mudah dalam penggunaan JST. Jadi salah satu kelemahan JST adalah penentuan arsitektur yang optimal, yang dimaksud arsitektur adalah penentuan struktur dan bobot-bobot koneksi dalam JST.

Evolutionary Algorithms (EAs) adalah algoritma-algoritma optimasi yang berbasis evolusi dalam dunia nyata. Oleh karena itu EAs dapat digunakan dalam optimasi pencarian arsitektur yang optimal dari JST. Algoritma EAs yang digunakan dalam tugas akhir ini adalah *Evolution Strategies* (ES). Pemilihan algoritma ES disebabkan kecepatan proses ES lebih baik dibandingkan dengan *Genetic Algorithm*[9].

Setiap atribut data memiliki pengaruh yang berbeda-beda dalam pengklasifikasian data. Hal ini tergantung pada seberapa besar nilai keinformatifan atau kontribusinya suatu atribut dalam pengklasifikasian data. Sehingga diperlukan *feature selection* terhadap atribut yang akan dijadikan sebagai input dalam pengklasfikasian data menggunakan JST.

Pada tugas akhir penulis mencoba menganalisis metode klasifikasi JST yang dipadukan dengan ES dan melakuan proses *feature selection* pada saat *preprocessing*. Metode JST yang digunakan adalah Feedforward Networks dengan *Supervised Learning*.

1.2 Perumusan Masalah

Dapat dirumuskan beberapa masalah yang dapat diangkat melalui penelitian Tugas Akhir ini, yaitu :

1. Bagaiman menerapkan proses *feature selection* dengan *information gain*.
2. Bagaimana membuat suatu model *classifier* dengan menerapkan metode JST dan ES?
3. Bagaimana menganalisa sistem klasifikasi berdasarkan parameter-paremeter masukkannya pada metode ES (seperti ukuran populasi struktur, ukuran populasi bobot serta ukuran *mutation step size*) untuk mendapatkan struktur dan bobot yang optimal pada JST.

Batasan masalah dari penelitian Tugas Akhir ini yaitu :

1. Jenis JST yang digunakan adalah *Feedforward Networks* dengan *Supervised Learning.*
2. Arsitektur JST yang digunakan adalah satu *hidden layer*.
3. Proses *discretization* menggunakan dua nilai diskret pada saat *preprocessing*.
4. Terdatapat dua data yang digunakan adalah data *Pima Indians Diabetes* dan Breast Cancer yang diambil dari www.ics.uci.edu [5]
5. Aplikasi yang dibangun bersifat *stand alone*.
6. Menggunakan MATLAB.

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah sebagai berikut :

1. Menerapkan *information gain* pada suatu data set.
2. Mengimplementasikan ES dalam menentukan struktur dan bobot-bobot yang optimal pada JST dalam membangun sebuah classifier.
3. Melakukan analisis terhadap performansi sistem dan parameter-parameter masukan ES untuk mendapatkan struktur dan bobot yang optimal pada JST.

1.4 Metodologi Penyelesaian Masalah

Metodologi yang digunakan dalam Tugas Akhir ini sebagai berikut :

1. Studi Literatur
   a. Pencarian referensi dan sumber-sumber yang berhubungan dengan JST, ES dan *Feature Selection*.
   b. Pendalaman materi JST, ES dan *Feature Selection*.
   c. Pengumpulan data dari www.ics.uci.edu
2. Perancangan Sistem
   Pada tahap ini dilakukan perancangan sistem yang akan diimplementasikan dengan menentukan parameter-parameter yang akan digunakan dalam pembangunan sistem.
3. Implementasi dan Training
   Pada tahap ini dilakukan implementasi ES dalam mengoptimasi dan melatih JST sesuai dengan perancangan sistem.
4. Testing dan Analisis Hasil
   Pada tahap ini dilakukan *testing* terhadap JST yang dihasilkan dari proses Implementasi dan Training, kemudian dilakukan analisis terhadapa hasil *testing* dalam mengklasifikasikan data.
5. Penyusunan Laporan
   Penyusunan laporan dalam bentuk buku Tugas Akhir dengan mengikuti kaidah penulisan yang berlaku dan berdasarkan hasil penelitian.

1.5 Sistematika Penulisan

Penulisan Tugas Akhir ini dibagi dalam lima bab, yang terdiri atas:

BAB 1 : Pendahuluan
Bab ini berisi latar belakang, rumusan masalah, batasan masalah, tujuan, metodologi penyelesaian masalah dan sistematika penulisan yang digunakan dalam penyusunan Tugas Akhir.

BAB II : Landasan Teori
Bab ini memuat dasar teori yang mendukung dan mendasari penulisan tugas akhir, yaitu mengenai *discretization*, *feature selection*, Jaringan Syaraf Tiruan dan *evolution strategies*.

BAB III : Analisis dan Perancangan Sistem
Bab ini menguraikan gambaran umum sistem serta pemodelan sistem.

BAB IV : Implementasi dan Pengujian Sistem
Bab ini berisi uraian mengenai implementasi dan analisis hasil percobaan yang telah dilakukan.

BAB V : Penutup
Bab terakhir ini menjelaskan kesimpulan umum dari seluruh rangkaian penelitian yang dilakukan dan saran untuk pengembangan selanjutnya.

5 Kesimpulan dan Saran

5.1 Kesimpulan

Dari hasil pengujian dan analisis dapat diambil kesimpulan sebagai berikut:

1. Kombinasi nilai parameter pada *Evolution Strategies* (ES) yang mampu menghasilkan JST yang optimal untuk klasifikasi data *Pima Indian Diabetes* ini yaitu ukuran populasi struktur 5, ukuran populasi bobot 50 dan nilai *mutation step size* 0.01 – 0.2 dan untuk data Breast Cancer ukuran struktur 5, ukuran populasi bobot 50 dan nilai mutation step size 0.01 – 0.8.
2. Dengan mengevaluasi ruang solusi yang terbatas, ES dapat menemukan solusi optimal pada kedua kasus data yang digunakan pada tugas akhir ini.
3. Arsitektur jaringan pada JST yang digunakan adalah 7 *input node*, 5 *hidden node* dan 35 koneksi antara *input* dan *hidden node* untuk data *Pima Indians Diabetes* dan 6 *input node*, 2 *hidden node* dan 12 koneksi untuk data *Breast Cancer*.
4. Sistem klasifikasi yang dihasilkan mempunyai performansi yang baik, hal ini ditunjukkan pada *fitness testing* sebesar 81,2834 % untuk data Pima Indians Diabetes dan 98,2456% untuk data Breast Cancer.

5.2 Saran

Beberapa saran yang dapat diberikan untuk pengembangan dan perbaikan di waktu yang akan dapat sebagai berikut :

1. Dapat menggunakan metode lain seperti *Support Vector Machine* (SVM) dan *Bayesian Network* untuk klasifikasi data.
2. Dapat menggunakan metode optimasi lain seperti *harmony search* atau *Evolution Algorithms* lain untuk struktur dan bobot JST.

Daftar Pustaka

[1] Achadin, Deviya Mutoharoh. 2010. *Analisis dan Implementasi Evolutionary Recurrent Neural Networks dalam Studi Kasus Peramalan Kurs Jual Emas*. Bandung. IT Telkom. Tugas Akhir

[2] Awarini, Pratiwi. 2008. Analisis dan Implementasi Data Mining Task Klasifikasi menggunakan Jaringan Syaraf Tiruan dengan *Feature Selection*. Bandung. IT Telkom. Tugas Akhir

[3] Data mining. Available at http://en.wikipedia.org/wiki/Data_mining. Diakses tanggal 17 juli 2010.

[4] Octaviani, Charisa Afia. 2010. *Analisis dan Implementasi Evolution Strategies Pada Kasus Prediksi Curah Hujan*. Bandung. IT Telkom. Tugas Akhir

[5] Pang-Ning Tan, Michael Steinbach and Vipin Kumar. 2005. Introduction to Data Mining

[6] *Pima Indians Diabetes.* Available at http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.

[7] Risvik,Knut Magne. *Discretization* of Numerical Attributes.

[8] Siang, Jong Jek, 2004, *Jaringan Syaraf Tiruan dan Pemrogramannya Menggunakan Matlab*, Jogjakarta, Andi.

[9] Smith,J.W., Everhart,J.E., Dickson,W.C., Knowler,W.C., & Johannes,R.S. 1988. *Using the ADAP learning algorithm to forecast the onset of diabetes mellitus*. IEEE Computer Society Press.

[10] Suyanto, 2007, *Artificial Intelligence Searching, Reasoning, Planning dan Learning*, Bandung, Informatika.

[11] Sykacek, Peter and Roberts,Stephen J. 2002. *Adaptive Classification by Variational Kalman Filtering*. NIPS

[12] Suyanto, 2008, *Soft Computing*, Bandung, Informatika.

[13] Wolberg, W.H. and Mangasarian, O.L. 1990. *Multisurface method of pattern separation for medical diagnosis applied to breast cytology*. In Proceedings of the National Academy of Sciences