

RULE MINING UNTUK KLASIFIKASI DATA MENGUNAKAN *SEARCH TREE*

Rina Dewi Indahsari¹, Broto Poernomo T.P²

^{1,2}STMIK Asia Malang

E-mail: ¹rideinsar30@gmail.com, ²papung@gmail.com

Abstrak – Penelitian ini adalah kajian teori dalam penerapan search tree untuk proses rule mining. Klasifikasi merupakan proses untuk menyatakan suatu objek ke salah satu kategori yang sudah didefinisikan sebelumnya. Search tree bekerja dengan pendekatan berbasis ruang solusi (State Space). Dalam kajian ini, search tree digunakan untuk menterjemahkan bobot hasil training dalam Jaringan Syaraf Tiruan (JST) menjadi set aturan konjungsi (IF-THEN). Proses ini disebut dengan rule mining. Dalam pembahasan akan dijelaskan proses training pada dataset yang cukup populer yaitu permasalahan “Playing Tennis”. Proses pelatihan (training) dilakukan dengan JST Metode Backpropagation. Dataset “playing tennis” memiliki 4 input atribut dan satu output atribut (atribut kelas). Untuk kepentingan rule mining, nantinya dataset akan dimodelkan menjadi 10 input atribut dan 2 output atribut. Formulasi permasalahan ke dalam jaringan syaraf dengan menggunakan 10 input neuron yang masing-masing mewakili value atribut non kelas dan dua output neuron yang mewakili value atribut kelas.

Kata Kunci — Klasifikasi, jaringan syaraf tiruan, backpropagation, search tree, ekstraksi rule

Abstract – This research is the study of theory in the application of the search tree for rule mining process. Classification is the process to declare an object into one of the categories that have been defined previously. Search tree-based approach to working with the solution space (State Space). In this study, the search tree is used to translate the results of weight training on Artificial Neural Network (ANN) into a set of rules conjunctions (IF-THEN). This process is called the rule mining. In the discussion will

explain the process of training on a dataset that is quite popular, namely the issue of “Playing Tennis”. The training process (training) conducted by JST Backpropagation method. Dataset “playing tennis” has 4 inputs and one output attribute (class attribute). For the purposes of rule mining, the dataset will be modeled to 10 inputs and 2 outputs attribute attribute. Formulation of problems into the neural network with 10 input neurons each each representing the attribute values of non-class and two output neurons that represent the attribute values of the class.

Keywords — Classification, artificial neural network, backpropagation, search tree, rule extraction

1. PENDAHULUAN

Rule mining merupakan kegiatan menambang aturan (rule) dari sekumpulan data yang kurang terstruktur. Aturan (rule) yang dihasilkan dari proses rule mining berbentuk aturan yang lebih terstruktur sehingga mudah dipahami manusia. Salah satu contoh data yang tidak familiar dengan manusia adalah hasil training jaringan syaraf tiruan, data hasil training ini dalam bentuk bobot (angka) yang susah dimengerti oleh manusia.

Jaringan syaraf tiruan merupakan alat yang baik dan telah digunakan untuk mengembangkan berbagai aplikasi pada dunia nyata, khususnya pada kasus-kasus dimana metode pemecahan masalah dengan cara tradisional gagal menyelesaikannya. Jaringan syaraf menunjukkan keuntungan seperti kemampuan belajar yang ideal dari sekumpulan data, kemampuan klasifikasi dan generalisasi untuk situasi tidak terstruktur, komputasi yang cepat meskipun melakukan

pemrosesan secara paralel dan toleransi terhadap noise. Keunggulan ini yang membuat jaringan syaraf sukses diterapkan untuk berbagai masalah dalam dunia nyata, termasuk: pengenalan suara, diagnosa medis, komputasi gambar, proses kontrol dan pemodelan diagnosis kesalahan. Namun para ahli tidak puas hanya dengan tingkat akurasi yang tinggi yang ditunjukkan oleh jaringan syaraf. Karena cara penalaran yang digunakan jaringan syaraf untuk mencari jawaban tidak dapat dilakukan dengan cepat, ini memberikan bukti perlunya untuk mengekstrasi pengetahuan yang terdapat pada data terlatih jaringan syaraf dan menunjukkannya secara simbolik untuk mendapatkan output-nya.

Salah satu pendekatan yang digunakan untuk ekstraksi aturan adalah metode berbasis pencarian ruang solusi. Metode ini terdiri dari 2 tahap, yang pertama adalah tahap pencarian aturan, dan tahap kedua adalah proses pengecekan validitas rule yang telah diperoleh dari tahap pertama. Pencarian aturan dimulai dari kemungkinan aturan dengan 1 premis sampai aturan dengan banyak premis. Jika aturan dengan 1 premis telah valid, maka tidak perlu dicari aturan dengan premis banyak. artikel menggunakan *font Times New Roman* berukuran 10 point dengan spasi satu. Penulis cukup mengetikkan (*me-replace*) kata atau kalimat yang ada di dalam template. Pendahuluan menguraikan latar belakang permasalahan yang diselesaikan, isu-isu yang terkait dengan masalah yg diselesaikan, ulasan penelitian yang pernah dilakukan sebelumnya oleh peneliti lain yang relevan dengan penelitian yang dilakukan.

2. KAJIAN TEORI

2.1 *Klasifikasi Dalam Data Mining*

Secara sederhana data mining merupakan proses ekstraksi informasi atau pola dari data yang ada dalam basis data yang besar. Data mining menjadi penting karena banyaknya data yang terkumpul saat ini, cepatnya transfer data yang terjadi pada saat ini serta adanya kebutuhan untuk dapat mengolah data mentah menjadi data yang bernilai dengan cepat dan tepat.

Klasifikasi merupakan salah satu kegiatan data mining untuk menyatakan suatu

objek ke salah satu kategori yang sudah didefinisikan sebelumnya. Klasifikasi adalah suatu proses untuk mengelompokkan sejumlah data ke dalam kelas-kelas tertentu yang sudah diberikan berdasarkan kesamaan sifat dan pola yang terdapat dalam data-data tersebut. Klasifikasi merupakan pencarian sekumpulan model (fungsi) yang menggambarkan dan membedakan kelas atau konsep data dengan maksud menggunakan model tersebut sebagai prediksi terhadap kelas atau objek dimana label kelas tersebut tidak diketahui. Beberapa metode klasifikasi pada data mining yang biasa digunakan adalah decision tree, bayesian, jaringan syaraf backpropagation (propagasi balik), k-nearest neighbor, Rule-based (berbasis aturan), Support Vector Machine (SVM) dan lain-lain.

2.2 *Backpropagation-Feed Forward Neural Network*

Neural Network (NN) atau Jaringan Syaraf Tiruan (JST) adalah prosesor tersebar paralel yang sangat besar (*massively parallel distributed processor*) yang memiliki kecenderungan untuk menyimpan pengetahuan yang bersifat pengalaman dan membuatnya siap untuk digunakan (Aleksander & Morton, 1990). Jaringan syaraf tiruan merupakan salah satu representasi buatan dari otak manusia yang selalu mencoba untuk mensimulasikan proses pembelajaran pada otak manusia tersebut.

Backpropagation merupakan algoritma pembelajaran yang terawasi (*supervised learning*) dan biasanya digunakan oleh JST dengan banyak lapisan untuk mengubah bobot-bobot yang terhubung dengan neuron-neuron yang ada pada lapisan tersembunyinya (*hidden layer*). JST backpropagation memiliki arsitektur jaringan yang *full-connected*. Dimana semua neuron input akan terhubung dengan semua neuron hidden, dan semua neuron hidden berhubungan dengan semua neuron output.

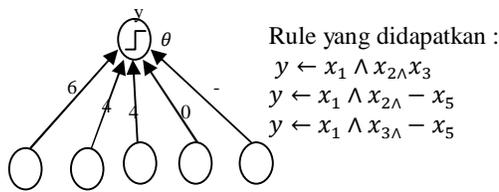
Algoritma backpropagation menggunakan error output untuk mengubah nilai bobot-bobotnya dalam arah mundur (*backward*). Untuk mendapatkan error ini, tahap perambatan maju (*forward propagation*) harus dikerjakan terlebih dahulu. Pada saat perambatan maju, neuron-neuron diaktifkan dengan menggunakan fungsi aktivasi sigmoid, yaitu:

$$f(x) = 1/(1 + e^{-x})$$

2.3 Rule Mining (Ekstraksi Rule) dari Jaringan Syaraf Tiruan

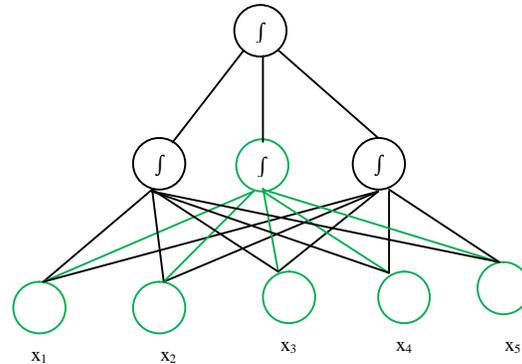
Ekstraksi aturan (rule extraction) dari jaringan syaraf tiruan merupakan suatu proses untuk memperoleh representasi yang tepat dari sebuah jaringan syaraf tiruan. Bentuk representasi yang terbentuk dapat berupa conjunctive (if-then) rules, m-of-n rules, fuzzy rules, decision tree, dan finite state automata

Proses ekstraksi aturan ini dilakukan setelah tahap pembelajaran dengan jaringan syaraf tiruan selesai dan menghasilkan bobot-bobot pelatihan. Bobot-bobot hasil pelatihan dari jaringan syaraf tiruan tersebut yang nantinya akan digunakan untuk proses ekstraksi rule/aturan.



Gambar 1. Hubungan antara arsitektur ekstraksi aturan dan rule yang didapatkan

Gambar 1 menunjukkan hubungan arsitektur ekstraksi dan rule yang dihasilkan serta mengilustrasikan proses ekstraksi rule dari sebuah jaringan syaraf tiruan yang sangat sederhana. Jaringan syaraf tiruan tersebut hanya terdiri dari lapisan input (terdiri dari 5 node input) dan lapisan output yang hanya memiliki 1 node. Nilai boolean true diwakili oleh nilai 1 dan nilai boolean false diwakili oleh nilai 0. Bentuk rule yang dihasilkan berbentuk conjunctive rule dan menggambarkan kondisi-kondisi input yang harus dipenuhi agar output bernilai 1. Sebagai contoh rule $y \leftarrow x_1 \wedge x_2 \wedge \neg x_5$ mengindikasikan bahwa agar y bernilai 1 (true) maka input x1, x2 harus bernilai 1 (true) dan x5 harus bernilai 0 (false).



Gambar 2. Arsitektur ekstraksi untuk JST dengan hidden layer

Untuk jaringan syaraf tiruan yang memiliki hidden layer seperti yang ditunjukkan oleh gambar 2 maka jaringan syaraf tiruan tersebut akan dipecah menjadi beberapa bagian. Bagian yang dihasilkan adalah jaringan syaraf tiruan yang terdiri dari input layer dan hidden layer dan jaringan syaraf tiruan yang terdiri dari output layer dan hidden layer.

3. PEMBAHASAN

Secara umum sistem terbagi menjadi 2 proses yaitu, proses pelatihan dan proses ekstraksi aturan. Proses pelatihan membutuhkan parameter JST yang diinputkan oleh user, serta dataset yang akan dilatihkan. Parameter JST yang diperlukan adalah jumlah neuron pada layer hidden, learning rate, maksimum iterasi dan target error yang diinginkan. Dataset yang akan dilatih telah tersimpan dalam database. Pelatihan dilakukan dengan jaringan syaraf tiruan backpropagation sampai target eror terpenuhi atau mencapai maksimum iterasi yang telah ditentukan.

Dari hasil pelatihan dihasilkan bobot-bobot yang kemudian digunakan sebagai inputan dalam proses ekstraksi aturan. Proses ekstraksi aturan dimulai dengan membentuk pohon pencarian dan kemudian dilakukan penelusuran pohon pencarian. Pohon pencarian yang dibentuk sebanyak kelas pada

output atribut dan ditelusuri sebanyak jumlah layer hidden.

3.1 Dataset Playing Tennis

Dataset untuk Playing Tennis terdiri dari 5 atribut yaitu 4 atribut non kelas (Outlook, Temperature, Humidity, Wind) dan 1 atribut kelas (PlayTennis). Atribut Outlook dan Temperature memiliki 3 distinct value. Sedangkan atribut Humidity dan Wind memiliki 2 distinct value. Demikian juga dengan atribut kelas PlayTennis juga memiliki 2 distinct value yaitu Yes dan No.

Tabel 1. Dataset Playing Tennis

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Dataset Playing Tennis tersebut harus dimodelkan terlebih dahulu untuk dapat melalui proses training dengan jaringan syaraf tiruan. Untuk memudahkan dalam menentukan atribut, maka setiap value atribut akan dinotasikan dengan angka integer 1, 2, dan seterusnya sebanyak jumlah value pada atribut yang dimaksud.

Dalam contoh kasus ini, dataset Playing Tennis akan diubah menjadi bentuk meta data seperti pada tabel 2. Keseluruhan formulasi

terdapat 12 value atribut dari dataset Playing Tennis. Terdiri dari 10 value input atribut non kelas dan 2 atribut kelas. Setiap value atribut akan dinotasikan dalam sebuah angka dan akan menjadi sebuah node/neuron input pada arsitektur jaringan syaraf tiruan. Masing-masing node tersebut akan memiliki 2 nilai distinct yaitu 0 dan 1. Pemodelan value atribut ke dalam angka nantinya akan memudahkan dalam penentuan notasi dalam pembentukan training set yang akan digunakan sebagai data inputan dalam proses pembelajaran dengan jaringan syaraf tiruan.

Tabel 2. Meta Data untuk Playing Tennis

No	Atribut	Value	Arti
1	Outlook	1	Sunny
2	Outlook	2	Overcast
3	Outlook	3	Rainy
4	Temperature	1	Hot
5	Temperature	2	Mild
6	Temperature	3	Cool
7	Humidity	1	High
8	Humidity	2	Normal
9	Windy	1	Weak
10	Windy	2	Strong
11	Play	1	Yes
12	Play	2	No

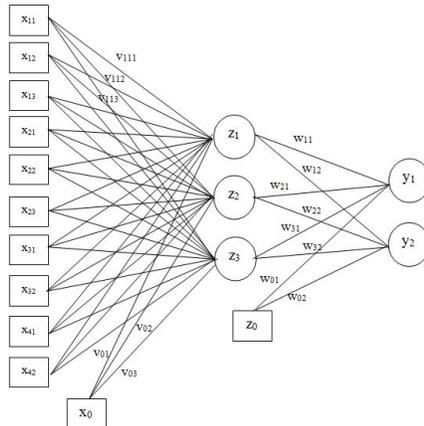
3.2 Formulasi Permasalahan Playing Tennis ke dalam JST Backpropagation

Tabel 3. Training Set Playing Tennis

x11	x12	x13	x21	x22	x23	x31	x32	x41	x42	y1	y2
1	0	0	1	0	0	1	0	0	1	0	1
1	0	0	1	0	0	1	0	1	0	0	1
0	1	0	1	0	0	1	0	0	1	1	0
0	0	1	0	1	0	1	0	0	1	1	0
0	0	1	0	0	1	0	1	0	1	1	0
0	0	1	0	0	1	0	1	1	1	0	1
0	1	0	0	0	1	0	1	1	0	1	0
1	0	0	0	1	0	1	0	0	1	0	1
1	0	0	0	0	1	0	1	0	1	1	0
0	0	1	0	1	0	0	1	0	1	1	0
1	0	0	0	1	0	0	1	1	0	1	0
0	1	0	0	1	0	1	0	1	0	1	0
0	1	0	1	0	0	0	1	0	1	1	0
0	0	1	0	1	0	1	0	1	0	0	1

Dataset Playing Tennis diubah menjadi training set dengan 12 atribut yaitu 10 atribut

non kelas ($x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}, x_{31}, x_{32}, x_{41}, x_{42}$) dan 2 atribut kelas (y_1, y_2). Masing-masing atribut memiliki 2 distinct value yaitu 0 dan 1. Training set untuk Playing Tennis ditunjukkan pada tabel 3.



Gambar 3. Arsitektur JST untuk Ekstraksi Rule Dataset Palying Tennis

Selanjutnya akan dilakukan training dengan JST. Untuk melakukan training dibutuhkan parameter jaringan syaraf tiruan yaitu jumlah hidden neuron, learning rate, target error dan maksimum epoch. Masing-masing parameter akan memiliki pengaruh yang berbeda terhadap proses pelatihan jaringan syaraf tiruan.

Training set Playing Tennis pada contoh ilustrasi ini akan diformulasikan ke dalam arsitektur jaringan syaraf tiruan dengan parameter sebagai berikut:

- Jumlah node hidden = 3
- Learning rate = 1
- Target error = 0.001
- Maksimum epoch = 5000

Penentuan notasi untuk pemodelan jaringan syaraf tiruan dataset Playing Tennis digunakan 10 input neuron dan 2 output neuron yang ditunjukkan pada gambar 3. Selanjutnya akan dilakukan proses pelatihan dan hasilnya ditunjukkan pada tabel dibawah ini.

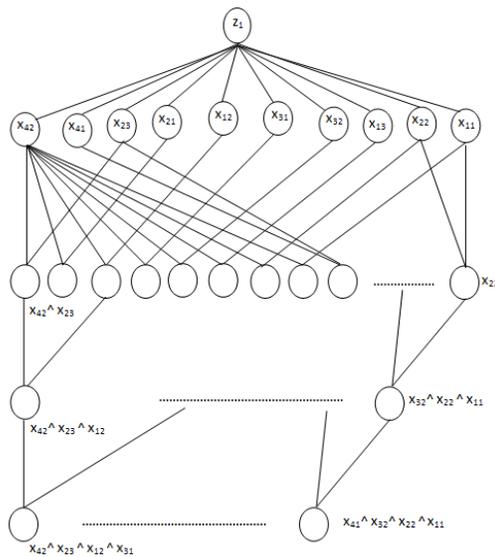
Bobot Hasil Pembelajaran Dataset Playing Tennis

Bobot Node Antara Input Layer dengan Node Hidden Layer			
	z_1	z_2	z_3
x_{11}	$v_{111} = 3.8424$	$v_{112} = -0.0765$	$v_{113} = -3.5544$
x_{12}	$v_{121} = -2.6206$	$v_{122} = -0.7554$	$v_{123} = 2.1273$
x_{13}	$v_{131} = 1.0894$	$v_{132} = -3.4211$	$v_{133} = -3.2172$
x_{21}	$v_{211} = -3.7632$	$v_{212} = 1.6487$	$v_{213} = 3.0054$
x_{22}	$v_{221} = 3.8329$	$v_{222} = 1.6808$	$v_{223} = -0.9568$
x_{23}	$v_{231} = 0.2695$	$v_{232} = -0.7385$	$v_{233} = -0.9805$
x_{31}	$v_{311} = -0.9823$	$v_{312} = 0.0836$	$v_{313} = 2.3829$
x_{32}	$v_{321} = 1.0605$	$v_{322} = -0.9359$	$v_{323} = -1.6496$
x_{41}	$v_{411} = 0.0174$	$v_{412} = -2.4536$	$v_{413} = -1.4092$
x_{42}	$v_{421} = -1.3683$	$v_{422} = 1.3572$	$v_{423} = 0.8115$
x_0	$v_{01} = -0.0485$	$v_{02} = -0.0485$	$v_{03} = 2.7227$
Bobot Node Antara Hidden Layer dengan Node Output Layer			
	y_1	y_2	
z_1	$w_{11} = -0.0485$	$w_{12} = 5.3667$	
z_2	$w_{21} = 3.6732$	$w_{22} = -4.0656$	
z_3	$w_{31} = 5.4507$	$w_{32} = -4.9312$	
z_0	$w_{01} = 0.1386$	$w_{02} = -0.526$	

3.3 Proses Ekstraksi Rule dengan Search Based untuk Dataset Playing Tennis

Pencarian rule dilakukan pada setiap kelas yang ada, dalam kasus ini terdapat dua kelas yaitu y_1 (PlayTennis=1) dan y_2 (PlayTennis=2). Untuk setiap kelas akan dilakukan pencarian rule pada setiap hidden node dengan menelusuri ruang kandidat rule yang muncul. Pencarian rule mengacu pada nilai aktivasi yang diharapkan dari z_1 ($= a_1$), z_2 ($= a_2$) dan z_3 ($= a_3$). Apabila bobot bernilai positif maka nilai aktivasi yang diharapkan adalah 1. Apabila bobot bernilai negatif maka nilai aktivasi yang diharapkan adalah 0.

Untuk tiap-tiap hidden node dicari rule dengan menelusuri ruang kandidat rule yang muncul. Dalam kasus ini ada 3 hidden node, sehingga akan terbentuk 3 pohon pencarian. Untuk y_1 (PlayTennis = 1), agar nilai aktivasi $y_1 \approx 1$ maka harus menentukan nilai aktivasi untuk masing-masing hidden node yang terhubung dengan y_1 . Pada tabel 4 hasil training jaringan syaraf tiruan didapatkan bahwa bobot antara z_1 dan y_1 (w_{11}) bernilai negatif maka nilai aktivasi yang diharapkan dari z_1 (a_1) adalah 0. Maka semua input node yang terhubung dengan z_1 diurutkan terlebih dahulu berdasarkan nilai bobotnya dimulai dari yang terkecil sampai terbesar (ascending). Selanjutnya dibuatlah sebuah ruang kandidat rule berbentuk tree sesuai seperti pada gambar 4.



Gambar 4. Pohon Pencarian Dataset Playing Tennis untuk z_1 ($y=1$)

Dari gambar 4 terlihat bahwa pohon pencarian yang dibangkitkan memiliki kedalaman 4 level. Hal ini disebabkan oleh jumlah atribut non kelas berjumlah 4 yaitu Outlook, Temperature, Humadity dan Windy. Pada masing-masing level akan memiliki jumlah node yang berbeda tergantung dari kombinasi node pada level atasnya. Masing-masing node pada pohon pencarian menghasilkan sebuah rule, kemudian rule tersebut akan di cek kevalidannya dengan menjumlahkan nilai bobot dari rule tersebut dengan nilai bobot terbesar dari atribut lain yang belum tercover (digunakan) oleh rule tersebut dan juga dijumlahkan dengan nilai bias dari hidden node yang sedang dilakukan proses pencarian rule. Dan hasil penjumlahan harus kurang dari 0 ($x < 0$).

Setelah pada masing-masing hidden node dicari rule maka rule-rule tersebut akan digabungkan. Namun sebelum itu dilakukan proses penghilangan rule yang sama. Proses penghilangan ini dilakukan dengan melihat kesamaan nilai aktivasi dari sebuah rule dengan rule yang lain. Dan juga dilakukan pemilihan rule yang optimum dengan menghitung akurasi dari masing-masing rule.

Dari hasil ekstraksi dataset Playing tennis maka dihasilkan rule-rule sebagai berikut :

a. IF (outlook=2) OR (humidity=2 AND outlook=1) OR (windy=2 AND temperature=2 AND outlook=3) OR (windy=2 AND temperature=2 AND humidity=2) THEN play=1

b. IF (outlook=1 AND humidity=1) OR (outlook=1 AND temperature=1) THEN play=2

Ekstraksi rule dengan pohon pencarian menghasilkan rule sejumlah output neuron. Pada contoh ekstraksi dataset Playing Tennis menghasilkan 2 buah rule. Rule dari setiap hidden layer akan digabungkan menjadi kesatuan rule dengan operator OR. Pada kesimpulannya, ekstraksi rule dataset Playing Tennis tidak menghasilkan 2 rule akan tetapi sebanyak premis yang dipisahkan oleh operator OR.

4. PENUTUP

Dari ilustrasi pada dataset Playing Tennis terbukti dapat menghasilkan rule yang valid. Dari proses ekstraksinya dapat diambil sebuah kesimpulan bahwa setiap permasalahan yang akan diekstraksi rule-nya harus dimodelkan dalam bentuk representasi biner. Data tidak dapat dimodelkan dengan representasi bipolar untuk formulasi ke dalam arsitektur jaringan syaraf tiruan. Sehingga apabila terdapat data dengan representasi bipolar, harus diekstraksi dengan metode yang lain, bukan menggunakan pohon pencarian.

DAFTAR PUSTAKA

- [1] Bertalya,. 2009. "Konsep Data Mining Klasifikasi: Pohon Keputusan", Universitas Gunadarma
- [2] Chrisna, Temmy, Bijaksana, Arif, Muntina, Eddy, 2005, "Ekstraksi Rule Dari Jaringan Syaraf Tiruan Menggunakan Metode Search Based Pada Data Mining", STT Telkom Bandung.
- [3] Dancey, Darren, 2008, "Tree Based Methods for Rule Extraction from Artificial Neural Networks", Manchester Metropolitan University, United Kingdom
- [4] Duch, Wlodzislaw, Kordos, Miroslaw, 2005, "Search-based Training for Logical Rule Extraction by Multilayer Perceptron", School of Computer Engineering Institute of Computer Science Nanyang Technological

*University The Silesian University of
Technology, Singapore*

- [5] Kusumadewi, Sri. 2003, "Artificial Intelligent (Teknik dan Aplikasinya)", Graha Ilmu, Yogyakarta
- [6] Mark W, Craven, 1995. "Using Neural Network For Data Mining", School of Computer Science Carnigie Mellon Univercity, Pittsburgh
- [7] Nayak, Richi, 1999, "Gyan: A Methodology for Rule Extraction from Artificial Neural Networks", Queenslan University of Technology Australia
- [8] Palade, Vasile, Neagu, Daniel-Ciprian, Patton, Ron J., 2001, "Interpretation of Trained Neural Networks by Rule Extraction", The University of Hull, Cottingham Road, United Kingdom

Seminar Nasional Inovasi Teknologi
UN PGRI Kediri, 22 Februari 2017

ISBN : 978-602-61393-0-6
e-ISSN : 2549-7952

Halaman ini sengaja dikosongkan