KLASIFIKASI POLA KONTEN E-MAIL MENGGUNAKAN JARINGAN SYARAF TIRUAN METODE BACKPROPAGATION UNTUK PENGECEKAN SPAM E-MAIL DENGAN DATA ACUAN DMC 2003

Enggar Pradipa [1], Elkaf Rahmawan, M.Kom [2]

Dosen Universitas Dian Nuswantoro [2], Mahasiswa Universitas Dian Nuswantoro [1]

e-mail : epradipa@gmail.com [1]

Abstract : *Determining an e-mail classification that included in Spam category can be done with seeing the e-mail environments. One of that environments is the content of an e-mail. The problem of this method E-mail classification rise when the word references quantity reach up to hundreds, thousands and even in infinite counts. Data mining toward the basis data that used in references for finding the pattern of Spam e-mail classification can be a solution for this problem. With using a Back Propagation Artificial Neural Networks, classification pattern from the Data Mining can be found and represented in the weights from each nodes in that Neural networks structure. With this classification pattern, we can get the e-mail classification that have a good reliability and accuracy.*

Key Words : Classification Pattern, Data Mining, Back Propagation Artificial Neural Networks.

Abstrak : *Penentuan klasifikasi e-mail yang termasuk dalam kategori Spam dapat dilakukan dengan melihat environment yang ada pada Suatu e-mail. Salah satunya dengan melihat konten dari e-mail tersebut. Permasalahan dalam klasifikasi e-mail dengan cara ini muncul ketika kata-kata yang bisa dijadikan acuan dalam pengklasifikasian tersebut mencapai jumlah ratusan, ribuan dan bisa tak terhingga. Data Mining terhadap basis data yang digunakan dalam acuan untuk menemukan pola klasifikasi e-mail spam dapat menjadi solusi dari permasalahan ini. Dengan menggunakan Jaringan Saraf Tiruan metode Back Propagation, pola klasifikasi dari data mining dapat ditemukan dengan representasi berupa bobot-bobot dari setiap node yang terdapat pada struktur jaringan saraf tersebut. Dengan Pola inilah, klasifikasi e-mail dapat dilakukan dengan tingkat akurasi yang dapat dipercaya.*

Kata Kunci : Pola klasifikasi, Data Mining, Jaringan Saraf Tiruan Back Propagation.

1. Pendahuluan

Penentuan klasifikasi e-mail dapat dipelajari dari berbagai *environment* yang ada dalam suatu e-mail tersebut. Salah satunya adalah isi konten dari e-mail itu. Penggunaan kata-kata dalam suatu e-mail dapat menjadi acuan bagi kita atau komputer untuk mengetahui apakah e-mail tersebut termasuk dalam klasifikasi e-mail spam atau tidak. Permasalahan muncul ketika jumlah kata-kata yang bisa dijadikan acuan dalam pengklasifikasian tersebut mencapai dalam jumlah ratusan hingga ribuan. Bisa dibayangkan berapa lamanya proses identifikasi untuk satu buah e-mail saja. Untuk itulah diperlukan adanya *Data Mining* terhadap basis data yang digunakan dalam acuan klasifikasi e-mail tersebut.

Data mining dapat dilakukan dengan berbagai teknik pemrosesan yang memiliki keunggulan tersendiri pada setiap tekniknya. Salah satu teknik itu adalah dengan menggunakan Jaringan Saraf Tiruan (JST). Dengan teknik ini, klasifikasi dari suatu data sebagai fungsionalitas dalam Data Mining dapat ditemukan. Selain itu, salah satu keunggulan dari JST adalah tidak membuat asumsi *apriori* dari suatu distribusi data dan berbagai bentuk interaksi dari faktor-faktor lain yang sebenarnya tidak berpengaruh terhadap data tersebut[1] . Hal inilah yang kemudian menjadi alasan penulis dalam memilih teknik JST untuk digunakan dalam proses penentuan pola klasifikasi e-mail diatas.

Jaringan Saraf Tiruan sendiri memiliki berbagai metode tersendiri dalam menghasilkan klasifikasi tersebut. Salah satu metode yang umum digunakan adalah metode Backpropagation (Propagasi Balik). Keunggulan dari metode ini adalah mampu mengenali dan mengekstraksi pola-pola yang berjumlah besar dan kompleks[2] . Hal ini sangat sesuai dengan apa yang dibutuhkan dalam proses data mining. Maka dari itu, dengan penggunaan Jaringan Saraf Tiruan dengan metode Backpropagation ini diharapkan Data Mining dari basis data konten e-mail mampu

bekerja optimal dan menghasilkan klasifikasi yang dapat dipercaya.

1.1 Perumusan Masalah

Dari ilustrasi tersebut diatas yang sudah disebutkan pada latar belakang, maka dapat dirumuskan beberapa permasalahan sebagai berikut :

1. Bagaimana menerapkan Jaringan Saraf Tiruan (JST) dalam data sets yang berisi data acuan konten e-mail. Dalam hal ini data sets yang digunakan adalah Data Mining Cup 2003.
2. Bagaimana pola klasifikasi data dalam jumlah yang besar dapat ditemukan dengan metode Backpropagation.

1.2 Batasan Masalah

Dalam penyusunan Tugas Akhir ini permasalahan dibatasi dengan beberapa hal, yaitu :

1. Perancangan Jaringan Saraf Tiruan untuk studi kasus Data Mining Cup 2003 (DMC 2003).
2. Metode JST yang digunakan adalah Backpropagation dengan pembelajaran terbimbing.
3. Proses pembelajaran Jaringan Saraf Tiruan akan dilakukan hingga kondisi nilai SSE < 0,1 .
4. Testing data yang digunakan untuk proses klasifikasi e-mail direpresentasikan melalui *offline file plain text*.

1.3 Tujuan

Tujuan dari penelitian Tugas Akhir ini adalah sebagai berikut :

1. Mengimplementasikan Jaringan Saraf Tiruan dalam proses Data Mining sehingga klasifikasi data dapat ditentukan dengan mudah.
2. Menerapkan Metode Backpropagation sebagai metode dalam JST untuk mendapatkan hasil yang akurat dalam menangani pola yang berjumlah besar dan kompleks.

1.4 Manfaat

Adapun manfaat yang dapat diperoleh dari penyusunan tugas akhir ini adalah :

1. Secara Teoritis

Hasil penelitian ini diharapkan dapat menambah pengetahuan dan informasi bagi peneliti sekaligus peneliti selanjutnya yang akan melakukan penelitian yang sama mengenai penggunaan Jaringan Saraf Tiruan dalam ekstraksi pola dari suatu kumpulan data.

2. Secara Praktis

Penelitin ini dapat dijadikan bahan pertimbangan :

a. Bagi Penulis

Menambah pengetahuan tentang pemanfaatan Jaringan Saraf Tiruan dalam suatu permasalahan yang nyata dan juga sebagai salah satu syarat kelulusan dalam menempuh pendidikan di Universitas Dian Nuswantoro Semarang.

b. Bagi Anti-Spam Software Developer

Diharapkan mampu menjadi referensi khusus dalam mengembangkan *engine* Anti-Spam software yang akan sangat bermanfaat untuk para pengguna fasilitas e-mail.

c. Bagi Universitas Dian Nuswantoro

Diharapkan berguna bagi segenap civitas akademika dalam hal menanggapi berbagai persoalan yang dihadapi di luar lingkungan kampus dan juga menambah pembendaharaan perpustakaan.

2. Tinjauan Pustaka

2.1 Spam

2.1.1 Pengertian Spam

Spam adalah penyalahgunaan perangkat elektronik untuk mengirimkan pesan secara terus menerus tanpa dikehendaki oleh si penerima. Tindakan ini biasa dikenal dengan istilah Spamming. Bentuk spam yang dikenal secara umum meliputi : E-mail spam, instant messages spam, Usenet newsgroup spam, Search Engine spam, Blog spam,Wiki spam, Spam iklan baris dan Social media spam [3] .

Spam di surat elektronik (E-mail) mulai menjadi masalah ketika internet telah dibuka untuk umum pada pertengahan 1990-an. Pertumbuhan yang pesat dari tahun ke tahun hingga saat ini telah menghasilkan spam 80% - 85% dari seluruh e-mail di dunia.

Berbagai cara dilakukan untuk menekan penyebaran e-mail spam. Baik dari segi hukum hingga dari segi teknis, dalam hal ini adalah adanya sistem tertentu yang berfungsi sebagai detektor spam.

2.1.2 Deteksi Spam

Secara idealnya, mail server seharusnya bisa mendeteksi dan menolak e-mail yang merupakan spam. Namun, tidak seperti manusia yang bisa mengenali spam e-mail secara instan, komputer memiliki kesulitan dalam mendeteksi spam dengan tepat tanpa perintah atau acuan data yang tepat untuk klasifikasinya. Hal inilah kemudian yang memicu munculnya berbagai metode dan algoritma dalam pendeteksian e-mail spam.

Ada dua metode utama yang digunakan dalam pendeteksian spam : Client Based dan Content Based. Client based detection menggunakan IP address, Hostname, alamat email atau bahkan DNS dari pengirim e-mail untuk dijadikan dasar dari penentuan Spam atau tidak[8] .
Dalam Content Based detection, deteksi dilakukan dengan mengidentifikasi isi dari e-mail tersebut. Spam e-mail biasanya memiliki kata-kata dan pola-pola tertentu yang hampir sama. Hal inilah yang kemudian dijadikan acuan dalam klasifikasi e-mail tersebut. Kendala paling utama yang dihadapi oleh metode ini adalah dengan begitu beragamnya jumlah string dan pola penyusunannya di dalam setiap e-mail spam yang beredar. Diperlukan adanya metode lain yang mampu menangani jumlah data yang besar dan kompleks agar penggunaan Content Based Detection ini dapat bekerja secara maksimal.

2.2 Data Mining

Perkembangan yang pesat di bidang pengumpulan data dan teknologi penyimpanan di berbagai bidang, menghasilkan basis data yang terlampau besar. Namun, data yang dikumpulkan jarang dilihat lagi, karena terlalu panjang, membosankan, dan tidak menarik. Seringkali, keputusan -yang katanya berdasarkan data- dibuat tidak lagi berdasarkan data, melainkan dari intuisi para pembuat keputusan. Sehingga,

lahirlah cabang ilmu penggalian data ini [6] .

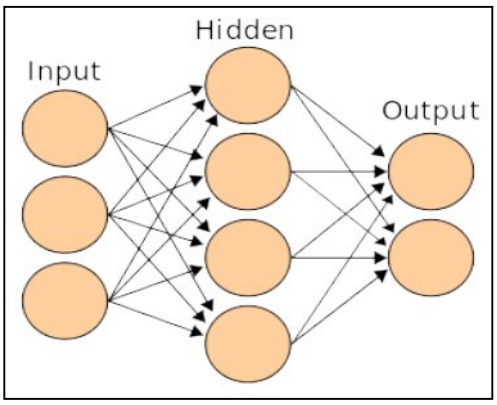
2.3 *Data Mining* adalah ekstraksi pola yang menarik dari data dalam jumlah besar. Suatu pola dikatakan menarik apabila pola tersebut tidak sepele, implisit, tidak diketahui sebelumnya, dan berguna. Pola yang disajikan haruslah mudah dipahami, berlaku untuk data yang akan diprediksi dengan derajat kepastian tertentu, berguna, dan baru. Penggalian data memiliki beberapa nama alternatif, meskipun definisi eksaknya berbeda, seperti KDD (knowledge discovery in database), analisis pola, arkeologi data, pemanenan informasi, dan intelegensia bisnis. Penggalian data diperlukan saat data yang tersedia terlalu banyak (misalnya data yang diperoleh dari sistem basis data perusahaan, e-commerce, data saham, dan data bioinformatika), tapi tidak tahu pola apa yang bisa didapatkan.

2.4 Jaringan Syaraf Tiruan

2.4.1 Pengertian Jaringan Saraf Tiruan

Jaringan Syaraf Tiruan dibuat pertama kali pada tahun 1943 oleh *neurophysiologist* Waren McCulloch dan *logician* Walter Pits, namun teknologi yang tersedia pada saat itu belum memungkinkan mereka berbuat lebih jauh.

Jaringan Syaraf Tiruan adalah paradigma pemrosesan suatu informasi yang terinspirasi oleh sistim sel syaraf biologi, sama seperti otak yang memproses suatu informasi. Elemen mendasar dari paradigma tersebut adalah struktur yang baru dari sistim pemrosesan informasi. Jaringan Syaraf Tiruan, seperti manusia, belajar dari suatu contoh. Jaringan Syaraf Tiruan dibentuk untuk memecahkan suatu masalah tertentu seperti pengenalan pola atau klasifikasi karena proses pembelajaran. Jaringan Syaraf Tiruan berkembang secara pesat pada beberapa tahun terakhir. Jaringan Syaraf Tiruan telah dikembangkan sebelum adanya suatu computer konvensional yang canggih dan terus berkembang walaupun pernah mengalami masa vakum selama beberapa tahun [4].



Gambar 2.1: Model dasar JST

2.4.2 Sum Square Error

Perhitungan kesalahan merupakan pengukuran bagaimana jarinngan dapatbelajar dengan baik sehingga dibandingkan dengan pola yang baru akan dengan mudah dikenali. Kesalahan pada keluaran jaringan merupakan selisih antara keluaran sebenarnya (current output) dan keluaran yang diinginkan (desired output). Selisih yang dihasilkan keduanya biasanya ditentukan dengan cara dihitung menggunakan suatu persamaan. Adapun rumus dari SSE adalah[7] :

$$\text{SSE} = \sum_p \sum_j (T_{jp} - X_{jp})^2$$

Dengan :

T_{jp} : Nilai keluaran Jaringan Saraf

X_{jp} : Nilai target.

2.4.3 Metode Backpropagation (Propagasi Balik).

Pembelajaran jaringan *backpropagation* (*backpropagation networks*) atau BPNN adalah satu tipe pembelajaran terbimbing (*supervised earning*) yang diterapkan untuk jaringan syaraf tiruan (*artificial neural networks*) dalam memecahkan suatu masalah.

Algoritma pembelajaran Backpropagation ini menggunakan metode *gradient descent*, yang menerapkan aturan pembaharuan bobot (*weight*) secara iteratif :

$$w_k \longleftarrow w_k - \mu \frac{\partial E}{\partial w_k}$$

Hingga bobot $w_k$dalam jaringan konvergen ke nilai (solusi) yang diharapkan. Pada persamaan (1), $\frac{\partial E}{\partial w_k}$adalah turunan parsial (*partial derivative*) error jaringan *E* terhadap bobot w_k, sedanghkan laju perubahan (modifikasi) bobot di setiap iterasi dikontrol oleh μ, yang disebut “laju pembelajaran” (*learning rate*).

3. Metodologi Penelitian

3.1 Analisa Data

Dalam menjalankan proses-proses didalamnya, JST yang akan dibangun memerlukan suatu kumpulan data yang mampu untuk dijadikan referensi belajar berupa kumpulan *string* yang sering digunakan dalam spam e-mail didalam kontennya. Dalam penelitian ini , datasets yang digunakan sebagai referensi adalah DMC 2003. Data yang ada pada DMC 2003 terdiri dari 8.000 data training dan 11.000 data testing yang masing-masing memiliki 543 atribut aktif yang bisa dipakai sebagai acuan dalam menentukan hasil output. Dari 8.000 data training, didapat klasifikasi dan distribusinya sebagai berikut :

1. Klasifikasi E-mail Spam berjumlah 3112 record (38,9% data training)
2. Klasifikasi E-mail non Spam berjumlah 4888 record (61,1% data training)

Data testing yang juga ada di dalam datasets ini akan digunakan sebagai data yang akan digunakan untuk men jaringan apakah hasil dari proses learning sudah cukup memadai untuk jaringan bisa dikatakan terlatih.

3.2 Gambaran Umum

Jaringan Saraf Tiruan yang akan dibuat diharapkan dapat menentukan klasifikasi suatu record masuk dalam kategori E-mail Spam atau tidak dengan representasi melalui Boolean character.

Alur proses yang terjadi adalah sebagai berikut :

1. Learning

   Learning dilakukan dengan menggunakan Data Training. Kondisi berhenti dari tahap learning adalah jika nilai SSE mencapai nilai lebih kecil dari 0,1. Hasil dari proses ini akan berupa sebuah model JST yang berisi nilai bobot-bobot di setiap vector yang ada pada jaringan beserta dengan biasnya. Setelah model jaringan didapatkan, dilakukan verifikasi model jaringan dengan menggunakan Data Testing dan disertakan struktur dan bobot JST ideal yang diperoleh dari hasil proses Learning. Bobot-bobot inilah

yang menjadi representasi dari pola klasifikasi.

2. Deteksi

   Proses ini dilakukan dengan menguji JST yang telah mengalami proses learning dengan sumber data *off-line plain text* sebagai representasi *on-line* e-mail. String yang kemudian dimasukkan ke dalam file adalah string didalam datasets DMC 2003.

3.3 Formulasi Permasalahan

Untuk JST Metode Backpropagation yang digunakan adalah :

1. Arsitektur JST yang dibangun terdiri tiga layer. Yaitu input layer, hidden layer dan output layer.
2. Pada input layer, vector inputan berasal dari atribut data yang aktif yaitu 543 atribut (ID dan Target tidak disertakan).
3. Pada hidden layer, hidden node tidak berkembang / tidak berevolusi.
4. Pada ouput layer, output node hanya terdapat 1 node. Dimana node tersebut mengeluarkan nilai diantara 0 atau 1.
5. Klasifikasi objek dilakukan dengan aturan sebagai berikut :
   - Apabila output bernilai = 1, maka objek akan diklasifikasikan sebagai kelas Spam.
   - Apabila output bernilai = 0, maka objek akan diklasifikasikan sebagai kelas Non Spam.
6. Struktur JST adalah *full connected* , yang artinya setiap input node memiliki koneksi dengan hidden node, dan setiap hidden node memiliki koneksi dengan output node.
7. Jumlah hidden node dihitung menggunakan rumus pada

$$Nh = \sqrt{Ni} * No$$

Dengan *Ni* adalah jumlah input node dan *No* adalah jumlah output node[9] . Dari rumus diatas maka didapatkan hidden node = $\sqrt{543} * 1 = 23$.

8. Inisialisasi bobot awal pada proses learning menggunakan metode inisialisasi Nguyen-Widrow.

Laju belajar / laju pelatihan (α) dinyatakan sebagai konstanta yang berharga antara 0,25 – 0,75.

4. Pengujian dan Hasil

4.1 Hasil Pelatihan

Sesudah algoritma beserta baris program lainnya di implementasikan, dimulailah tahap pelatihan dengan memasukkan nilai input kedalam jaringan saraf tiruan. Untuk mempermudah pengoperasian, penulis menggunakan Graphic User Interface (GUI) seperti yang ditampilkan pada gambar dibawah

. Hasil yang diharapkan keluar dari tahap pelatihan yang dilakukan ini adalah model Jaringan Saraf Tiruan yang terlatih dan memiliki nilai Sum Square Error (SSE) yang lebih kecil dari nilai 0,1[9] . Dari bentuk atau model dari JST Backpropagation yang sudah terlatih itulah didapatkan nilai-nilai bobot antara konektor input-hidden , bias hidden, konektor hidden-output dan bias output dari jaringan.

Dari kumpulan bobot-bobot yang telah didapatkan dari proses pelatihan inilah yang secara sekaligus menjadi pola klasifikasi *e-mail* untuk tahap pelatihan.

Gambar 4.2 : Tampilan GUI saat proses learning.

Dari pelatihan yang telah dilakukan dan diperlihatkan pada gambar diatas , secara umum informasi yang bisa didapatkan adalah :

- Pelatihan berhenti setelah melakukan *looping* Data Training pada generasi ke-7. Dengan kata lain jaringan telah melakukan proses pengubahan nilai bobot sebanyak 56.000 kali.
- Output yang sesuai dengan data target sebanyak 7.915 data dari 8.000 data training yang disediakan. Dengan kata lain, tingkat

akurasi kesamaan hasil output mencapai 98,9375 %.

- Nilai error sebanyak 85 data atau sebesar 1,0625 %.
- Nilai SSE final yang didapat adalah 0.0993402977039908.

Setelah mendapatkan hasil dari proses learning yang berupa bobot-bobot konektor dan bias jaringan, model jaringan tersebut perlu disimpan untuk bisa digunakan dan diproses selanjutnya. Pada penelitian ini, model jaringan akan disimpan ke dalam file JST.bin . File inilah yang nantinya akan digunakan saat awal dari proses verifikasi model jaringan dan deteksi e-mail.

5.1 Kesimpulan

1. JST dengan metode Backpropagation mampu untuk melakukan data mining untuk mendapatkan klasifikasi jenis e-mail yang sesuai dengan datasets yang diberikan dengan tingkat akurasi yang tinggi.
2. Kata-kata / String yang tidak dimiliki dalam referensi tidak akan dikenali dan akan menghasilkan klasifikasi yang tidak tepat.

5.2 Saran

1. Perlu adanya database yang mampu mencakup lebih banyak pola kata dan jenis bahasa agar JST yang dirancang mampu menjadi *engine spam detector* yang lebih bisa digunakan oleh setiap orang dari berbagai Negara

Perlu adanya pengembangan ke dalam bentuk program mandiri sebagai program anti spam yang mampu untuk menerima dan memproses data e-mail langsung secara online.

DAFTAR PUSTAKA

[1] Lisboa, P.J.G. and Vellido A. (2000) . Business Applications Of Neural Networks : The-State-of-the-art-of Real World Applications. Singapore : World Scientific Publishing.

[2] Puspitaningrum, Diyah. (2006) . Pengantar Jaringan Saraf Tiruan. Jogjakarta : Penerbit Andi.

[3] http://en.wikipedia.org/wiki/Spam_%28electronic%29, diakses tanggal 12 Februari 2013

[4] Yani, Eli. (2005) . Pengantar Jaringan Saraf Tiruan. Artikel Kuliah. http://trirezqiariantoro.files.wordpress.com/2007/05/jaringan_syaraf_tiruan.pdf, diakses tanggal 12 Februari 2013

[5] Nemati, Hamid R. and Barko, Christopher D. (2005) . Organizational Data Mining. USA : IGI Global.

[6] http://en.wikipedia.org/wiki/Data_mining, diakses tanggal 9 Februari 2013.

[7] Hermawan, Arief. (2006) . Jaringan Saraf Tiruan : Teori dan Aplikasi. Jogjakarta : Penerbit Andi.

[8] http://www.seaglass.com/postfix/spam-detection.html, diakses tanggal 11 Maret 2013.

[9] Suyanto, Algoritma Genetika Dalam Matlab. (2005). Jogjakarta : Penerbit Andi.