Prosiding Seminar Nasional Penelitian, Pendidikan dan Penerapan MIPA
Fakultas MIPA, Universitas Negeri Yogyakarta, 16 Mei 2009

PREDIKSI DATA HILANG MENGGUNAKAN NEURAL NETWORK

Winita Sulandari dan Sarngadi Palgunadi Yohanes

Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Sebelas Maret, Surakarta

Abstrak

Fakta menunjukkan bahwa tidak semua data di lapangan merupakan data yang lengkap, sering kita jumpai adanya data yang hilang (missing data). Dalam penelitian ini, neural network (NN) dipilih sebagai metode untuk prediksi data hilang. NN adalah suatu sistem proses informasi yang mempunyai karakteristik tampilan seperti pada jaringan syaraf biologis. Dalam penerapannya NN mengandung sejumlah parameter (bobot) yang terbatas. Jumlah parameter yang optimal tergantung pada penentuan kombinasi yang tepat antara jumlah variabel input dan jumlah unit pada lapisan hidden. Untuk menentukan jumlah unit pada lapisan hidden didasarkan pada kriteria informasi MSE.

Data yang digunakan berupa data lengkap, IHK periode Januari 2002 sampai dengan Januari 2007 (terdapat sebanyak 61 data). Data dieliminasi secara random sebanyak 5%, 10%, 15% dan 20% sebagai data hilang. Untuk prediksi data hilang dengan NN dilakukan pelatihan dari sebanyak 55 data pertama dikurangi banyaknya data hilang sehingga diperoleh model dengan bobot-bobot tertentu. Model ini selanjutnya akan digunakan untuk prediksi data hilang. Dari 5 data terakhir akan diketahui tingkat kebenaran NN dalam prediksi data IHK.

Lebih lanjut, penulis membandingkan metode NN dengan metode yang lain, yaitu substitusi mean dan mean dua data terdekat. Kesimpulan yang diperoleh adalah metode NN memberikan MSE paling kecil dibandingkan dengan metode lain (mean dan mean dua data terdekat), dalam hal prediksi 5 data terakhir dari data IHK yang digunakan.

PENDAHULUAN

Peramalan atau membuat prediksi merupakan unsur yang penting dalam pengambilan keputusan. Hal ini dikarenakan efektif atau tidaknya suatu keputusan umumnya tergantung pada beberapa faktor yang tidak dapat dilihat pada waktu keputusan diambil. Fakta menunjukkan bahwa tidak semua data dilapangan merupakan data yang lengkap, sering kita jumpai adanya data yang hilang (*missing data*). Beberapa tindakan dapat diambil untuk mengatasi kasus data hilang. Cara yang paling mudah adalah dengan membuang data hilang tersebut. Sayangnya cara ini dapat mengakibatkan hilangnya informasi yang berharga karena cacah data yang berkurang. Yohanes SP (1998) telah mengkaji beberapa pendekatan alternatif bagi estimasi nilai data hilang pada analisis diskriminan. Dalam tulisannya Yohanes SP mengungkapkan bahwa beberapa peneliti lain menggunakan algoritma *expectation-maximization* (EM) untuk estimasi data hilang. Algoritma ini didasarkan pada penggantian data hilang dengan estimasi *mean* dan *standart deviasi* populasi yang diiterasi sehingga mencapai konvergen.

Peneliti lain seperti Velicer dan Colby (2005) telah membandingkan empat metode untuk mengatasi adanya data hilang pada data runtun waktu. Keempat metode itu adalah (a) membuang data hilang, (b) substitusi mean, (c) substitusi dua data terdekat dan (d) *maximum likelihood* (ML). Dalam percobaan estimasi data hilang pada data AR(1) ini diperoleh kesimpulan bahwa ML menghasilkan estimasi yang paling akurat dibanding ketiga metode yang lain, sementara substitusi mean adalah sebaliknya.

Pada penelitian ini penulis akan mencoba memprediksi data hilang menggunakan Neural Network (NN). Untuk mengetahui tingkat kebenaran dari metode ini, penulis membandingkan dengan metode prediksi data hilang lain yaitu substitusi mean dan mean dua data terdekat. Data hilang yang dimaksud adalah bagian dari data runtun waktu. Dari beberapa penelitian yang

berkaitan dengan NN, menunjukkan bahwa NN telah berkembang sebagai alat prediksi data runtun waktu. Lapedes dan Farber merupakan salah satu dari beberapa peneliti pertama yang menggunakan NN untuk prediksi data runtun waktu. Selanjutnya banyak penelitian dilakukan berkaitan dengan prediksi pada data runtun waktu riil; antara lain dapat dilihat pada Faraway dan Chatfield (1998), Situngkir dan Surya (2003), Pavelka (2002) dan Heravi, Osborn, dan Birchenhall (2003). Dari beberapa penelitian tersebut, data yang digunakan adalah data yang lengkap. Sejauh ini, penulis pernah melakukan penelitian sebelumnya mengenai prediksi data emas menggunakan NN dengan mengabaikan adanya data hilang.

Agar pembahasan tidak meluas, penulis dalam penelitian ini membatasi: (1) NN terdiri dari dua lapisan, dengan fungsi tangen sigmoid sebagai fungsi aktivasi pada lapisan pertama dan fungsi identitas sebagai fungsi aktivasi pada lapisan kedua. (2) algoritma pelatihan yang digunakan adalah algoritma propagasi balik dengan 100 *epoch* pelatihan; (3) data finansial yang digunakan adalah indeks harga konsumen periode waktu Januari 2002 – Januari 2007; dan (4) alat penelitian yang digunakan adalah MATLAB.

METODE PENELITIAN

Metode Penelitian yang digunakan adalah studi literatur yang selanjutnya diterapkan pada kasus data finansial, yaitu data IHK bahan makanan di Indonesia periode Januari 2002 sampai dengan Januari 2007. Data IHK yang digunakan merupakan data lengkap. Adapun langkah-langkah yang diambil untuk menyelesaikan masalah disajikan sebagai berikut.

a. Data dibagi menjadi dua bagian, data pelatihan dan data uji (5 data terakhir)
b. Eliminasi data hilang sebanyak 5%, 10%, 15% dan 20% dengan bantuan software Microsoft Excel, dengan perintah RAND()
c. Prediksi data hilang menggunakan metode substitusi mean dan metode mean dua data terdekat dan dilanjutkan dengan penentuan model jaringan NN menggunakan metode *trial and error* untuk mendapatkan model terbaik. Kriteria informasi yang digunakan dalam pemilihan model terbaik adalah MSE
d. Prediksi data hilang menggunakan NN. Sebelumnya ditentukan terlebih dahulu model NN berdasarkan sisa data setelah proses eliminasi.
e. Prediksi 5 data terakhir menggunakan model yang diperoleh dari masing-masing metode (substitusi mean, mean dua data terdekat dan NN), dan membandingkan MSEnya

HASIL DAN PEMBAHASAN

Spesifikasi Data

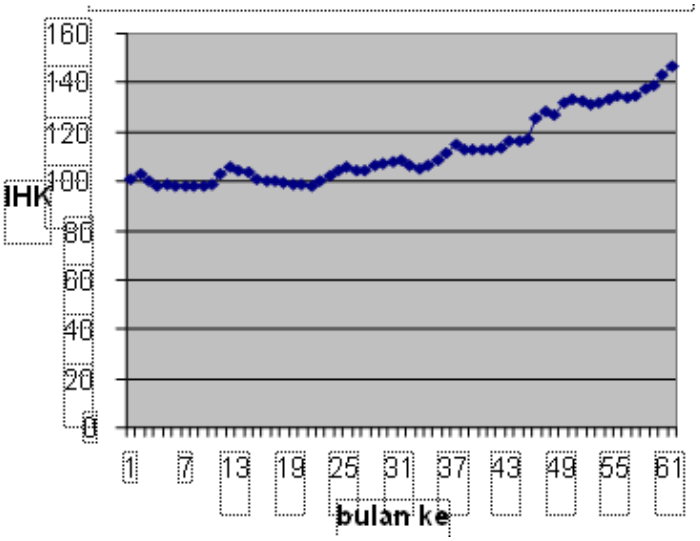
Data yang digunakan dalam penelitian ini adalah berupa data lengkap dari data IHK bahan makanan di Indonesia periode Januari 2002 sampai dengan Januari 2007 (terdapat sebanyak 61 data). Djarwanto (1989) mendefinisikan angka indeks sebagai suatu ukuran statistik yang menunjukkan perubahan suatu variabel atau sekumpulan variabel yang berhubungan satu sama lain pada waktu atau tempat yang sama atau berlainan. Secara umum indeks harga dihitung dengan rumus

$$I_n = \frac{\sum_i P_{ni}}{\sum_i P_{0i}} . 100$$

dengan

I_n : indeks harga pada periode ke-n,
P_{ni} : harga pada periode ke-n, satuan waktu dapat berupa tahun, bulan, minggu atau hari,
P_{0i} : harga pada tahun dasar.

Tahun 2002 diambil sebagai tahun dasar (sebagai dasar perbandingan) dikarenakan keadaan perekonomian pada tahun tersebut dianggap relatif stabil, dan tidak terlalu jauh dari tahun-tahun yang hendak dicari angka indeksnya. Angka indeks pada tahun dasar dinyatakan dengan 100, karena dianggap bahwa sesuatu hal pada tahun 2002 sebagai 100%.



Gambar 2. Data IHK Periode Januari 2002 – Januari 2007
(Sumber: www.bi.go.id)

Data *Prepocessing*

Prepocessing data bertujuan untuk efisiensi pada pelatihan dalam NN. Sebelum melakukan pelatihan, data input dan target ditransformasikan sehingga nilainya terletak pada interval tertentu, sesuai kebutuhan. Pada penelitian ini penulis menggunakan fasilitas yang ada pada MATLAB, yaitu fungsi **premnmax**. Dengan fungsi ini diperoleh

$$p_n = \frac{2(p - \min(p))}{\max(p) - \min(p)} - 1$$

dengan p_n = data hasil transformasi
p = data asli.

Dengan demikian data hasil transformasi dengan fungsi **premnmax** menghasilkan interval data [-1, 1]. Hal ini terkait dengan pemilihan fungsi aktivasi pada jaringan.

Proses Pelatihan Data

Pada awalnya dilakukan eliminasi data sebanyak yang diinginkan, mulai dari 5% hingga 20%. Setelah itu data hilang diprediksi dengan metode substitusi mean, mean dua data terdekat dan yang terakhir adalah dengan NN. Data yang ada ditransformasi menggunakan fungsi **premnmax** terlebih dahulu baru kemudian dibagi menjadi dua bagian, 55 data pertama sebagai data pelatihan dan 5 data terakhir sebagai data uji. Cara pelatihan jaringan pada kasus metode mean dan mean dua data terdekat sebagai pengganti data hilang adalah sama. Dipilih input dan target dari ke 55 data pertama. Misalkan diambil data pertama sebagai input maka data kedua sebagai target dan seterusnya. Untuk kasus data hilang misalkan data hilang adalah data ketiga, data kedua tidak dapat dijadikan input pada proses pelatihan. Hal ini disebabkan data ketiga tidak ada, sehingga tidak ada target. Jadi data input selanjutnya adalah data keempat dan data kelima sebagai target. Pelatihan menggunakan fungsi **newelm** dengan 1 unit input, 8 unit pada lapisan *hidden* dan 1 unit output.
Setelah model diperoleh, model akan digunakan untuk prediksi data hilang. Untuk kasus di atas, jika data ketiga adalah data hilang maka data kedua sebagai input. Input ini dimasukkan ke dalam model sehingga diperoleh suatu output sebagai hasil prediksi data ketiga. Proses ini berlangsung hingga semua data hilang terprediksi.

Keseluruhan data (termasuk hasil prediksi data hilang) dilatih lagi dengan cara yang sama ketika mencari model untuk prediksi data hilang, hanya saja model ini akan digunakan untuk prediksi 5 data terakhir data IHK.

Hasil Simulasi

Hasil simulasi dari prediksi 5 data terakhir data IHK disajikan pada tabel di bawah ini.

Prosentase data hilang	MSE		
	Metode mean	Metode mean dua data terdekat	Metode NN
5%	43.7747	32.9855	8.6831

10%	40.7321	55.2788	4.6996
15%	58.3963	42.2326	10.7693
20%	119.1764	8.9780	1.5870

Tabel di atas menunjukkan bahwa secara umum, untuk keempat kasus (prosentase data hilang 5%, 10%, 15% dan 20%), metode NN memberikan nilai MSE terkecil. Naik turun MSE pada metode mean dua data terdekat terlihat sebanding dengan metode NN sementara MSE pada metode mean tidak. Pada kedua metode, NN dan mean dua data terdekat, nilai MSE tidak memberikan arti bahwa semakin banyak data hilang semakin besar pula MSEnya.

KESIMPULAN

Berdasarkan beberapa percobaan yang dilakukan, model NN yang paling baik untuk prediksi data hilang adalah Elman network dengan 1 unit input, 8 unit hidden dan 1 unit outtput. Secara umum dapat disimpulkan bahwa metode NN memberikan hasil terbaik dibandingkan metode substitusi mean dan substitusi mean dua data terdekat jika dilihat dari MSE hasil prediksi 5 data terakhir data IHK.

DAFTAR PUSTAKA

Bishop, C. M. 1995. Neural Networks for Pattern Recognition. Oxford University Press. New York.

Connor J., Atlas L. E., & Martin D. R.. 1992. Recurrent Networks and NARMA Modelling, in Moody J. E., dkk.(eds.): *Neural Information Processing Systems 4*, Morgan Kaufmann, San Matteo, CA, pp.301 – 308.

_____. 1994. Recurrent Neural Networks and Robust Time Series Prediction. *IEEE Transactions On Neural Networks*, vol. 5, no. 2, pp. 240-253.

Demuth, H., & Beale, M. 1998. *Neural Network Toolbox User's Guide*. Math Works, Inc. Natick, MA.

Djarwanto, P.S. 1989. *Statistik Sosial Ekonomi*. BPFE Yogyakarta.

Dorffner, G. 2004. Neural Networks for Time Series Processing. http: //www.neci.nec.com/~lawrence/papers.html.

Faraway, J., & Chatfield, C. 1998. Time Series Forecasting with Neural Networks: a comparative Study Using The Airline Data. Royal Statistical Society. Appl., Statist., vol. 47, Part 2, pp. 231 – 250.

Fausett, L. 1994. Fundamentals of Neural Networks: Architectures, Algorithms, and Applications. Prentice-Hall. New Jersey.

Heravi, Saeed., Osborn, D. R., & Birchenhall, C. R. 2003. Linear versus Neural Network Forecast for European Industrial Production Series. www.ses.man.ac.uk/osborn/ **neural**%20**network**%20**forecast**s.pdf

Pavelka, A. 2002. Application of Autoregressive Models and Artificial Neural Networks in Time series Prediction. Institute of Chemichal Technology Prague.

Saputro dan Sulandari. 2005. Jaringan Syaraf Tiruan dengan Variasi Backpropagasi untuk Memprediksi Indeks Harga Emas. Laporan DIPA S1. Jurusan Matematika FMIPA UNS.

Siang Jong Jek. (2005). *Jaringan Syaraf Tiruan dan Pemrograman Menggunakan Matlab*. Cetakan pertama, Andi Offset, Yogyakarta.

Situngkir, H., & Surya, Y. 2003a. Keuangan Komputasional: Jaringan Saraf Buatan untuk Prediksi Data Deret Waktu Keuangan. Working Paper WPE2003. Bandung FE institute.

Sulandari dan Subanar. 2005. Neural Network Model ARMA untuk Prediksi Data Finansial. Sains dan Sibernatika. Berkala Penelitian Pascasarjana Ilmu-ilmu Sains Universitas Gadjah Mada. Vol. 18, no. 2.

Velicer, W.F., & Colby, S. M. 2005. A Comparison of Missing Data Procedures for ARIMA Time Series Analysis. Manuscript Educational and Psychological measurement publication.

Yohanes, S P. 1998. Missing Data Pada Analisis Diskriminan. Laporan Penelitian , Laboratorium Pusat MIPA UNS.