ANALISIS 8 (1 & 2), 101-109 (2001)

# PRAPEMPROSESAN DATA MENGGUNAKAN TEKNIK TAPISAN DALAM PEMODELAN PERCEPTRON MULTI ARAS

ROSHIDI DIN
KU RUHANA KU MAHAMUD
*Sekolah Teknologi Maklumat*
*Universiti Utara Malaysia*

## ABSTRAK

*Tujuan kajian ini dilakukan adalah untuk melihat kesan teknik tapisan terhadap prestasi model harga rumah dalam membuat peramalan berdasarkan penggunaan kaedah prapemprosesan data dalam kebolehupayaan pembelajaran rangkaian neural jenis perceptron multi aras (Multilayer Perceptron). Kajian ini juga mempertimbangkan perbandingan model pengujian data rangkaian neural dengan analisis regresi berganda. Data yang dilatih tidak hanya berasaskan kepada indeks harga rumah tetapi juga mengambilkira beberapa aspek lain yang berkaitan secara langsung dan tidak langsung dengan harga rumah. Dapatan daripada model kajian ini menunjukkan model harga rumah yang didapati menghasilkan ramalan yang lebih baik apabila teknik tapisan digunakan.*

## ABSTRACT

*This purpose of this research is to study the effect of filtering technique on house price modeling performance in predicting of learning capability of Multilayer Perceptron. This study also considered the comparison between neural network data testing modeling and multiple regression. Data is not only based on house price index but also includes other aspects which are related directly or indirectly to the house price index. From the finding, it is learnt that by using filtering technique, better prediction in house price modeling was obtained.*

## PENGENALAN

Prapemprosesan data adalah operasi yang pertama terhadap mana-mana set data (Famili *et al.,* 1997) dan sekurang-kurangnya perlu dilakukan sekali ke atas data dalam aplikasi dunia nyata. Sebab utama prapemprosesan ke atas data penting ialah untuk mencegah terlebih padanan (*overfitting*), mengelak kekakuan (*stiffness*) dalam proses pembelajaran dan memudahkan masalah difahami (Peterson *et al,* 1994). Ianya juga penting kerana sifat data dan pengetahuan bermakna daripada set data dapat dianalisis (Famili *et al.,* 1997). Tujuan utama analisis data adalah untuk menemui atribut atau ciri-ciri yang berkaitan dengan deskripsi data atau ramalan data dan menapis keluar atribut yang tidak berkaitan (Ivo & Gunter, 1998). Secara umum, terdapat dua kaedah prapemprosesan data yang biasa digunakan iaitu tranformasi dan penormalan (Lou, 1993). Tranformasi termasuklah memanipulasi input data mentah untuk mewujudkan input tunggal kepada rangkaian neural buatan manakala penormalan adalah tranformasi yang dilakukan ke atas input data tunggal lalu data diagihkan secara setara dan diskalakan dalam julat yang boleh diterima dalam rangkaian. Salah satu kaedah penormalan data adalah data diskalakan ke dalam julat di antara -1 dan 1 atau di antara 0 dan 1 (Peterson *et al.,* 1994).

Ramalan dalam rangkaian neural akan lebih tepat setelah data input yang dipilih dilakukan prapemprosesan terlebih dahulu (Lou, 1993) dan boleh menghasilkan pembelajaran yang cepat dengan set data yang kecil (Famili *et al.,* 1997). Didapati juga, ramalan dalam rangkaian neural jenis perceptron multi aras mudah menghasilkan penyelesaian yang baik dalam masa latihan yang lebih pendek dengan input tempoh masa yang lambat (Ohlsson *et al.,* 1994) dan ianya adalah lebih baik daripada *Learning Vector quantization* (LVQ) dalam semua keadaan fizik yang berkuasa tinggi apabila menghadapi masalah (Peterson *et al.,* 1994) serta luas penggunaannya kerana mudah dan berprestasi tinggi. Secara teorinya, rangkaian perceptron multi aras dengan satu lapisan tersembunyi sudah cukup untuk memodelkan mana-mana fungsi selanjar (Cybenko, 1989) manakala secara praktikalnya, dengan dua lapisan tersembunyi akan lebih efisien tetapi sukar untuk dilatih (Hartman & Keeler, 1991; Lönnblad *et al.,* 1992; Ohlsson *et al.,* 1994) dalam membuat ramalan. Secara kesimpulannya rangkaian perceptron multi aras dengan satu lapisan tersembunyi adalah cukup sesuai untuk pelbagai pengelasan kerja manakala

penggunaan dua lapisan tersembunyi dalam perceptron multi aras lebih sesuai untuk masalah pemadanan fungsi (Peterson *et al.*, 1994).

Aplikasi-aplikasi rangkaian neural yang telah digunakan dalam membuat ramalan ialah ramalan syarikat-syarikat muflis di pasaran (Marcus & Ramesh, 1990), ramalan harga bon (Soumitra & Sashi, 1998), ramalan harga stok harian (Herbert, 1998), ramalan indeks komposit bursa saham (Bahrom & Shang, 1996), ramalan penganggaran kos (Bode, 1998), ramalan kadar penempatan bilik (Law, 1998) dan simulasi gelagat pasaran harga rumah persendirian (Wang & Ho, 1995).

Objektif kajian ini adalah untuk membentuk model rangkaian neural yang melalui proses jujukan, rawak dan tapisan dalam prapemposesan data. Model kajian yang digunakan didasarkan kepada satu kajian yang menggunakan pendekatan rangkaian neural yang berasaskan perceptron multi aras untuk ramalan harga rumah (Ku-Mahamud *et al.*, 1999) yang hasilnya didapati bahawa prestasi ramalan menggunakan rangkaian neural adalah lebih baik daripada prestasi ramalan menggunakan analisis regrasi berganda walaupun data yang digunakan untuk membentuk model rangkaian neural tidak melalui proses tapisan. Data sebenar harga rumah dimasukkan ke dalam model rangkaian neural sebagai pengalaman pembelajaran yang terdiri daripada faktor-faktor yang mempengaruhi harga sesebuah rumah.

Seterusnya model ini diuji dengan data sebenar yang dipilih secara rawak daripada sampel untuk melihat ketepatan ramalan ke atas harga sesebuah rumah. Kemudian keputusan dari model tersebut akan dibandingkan dengan keputusan dari kaedah statistik dengan menggunakan data yang sama. Perbandingan ini bertujuan untuk membuktikan bahawa pendekatan yang digunakan dalam rangkaian neural dengan teknik tapisan akan menghasilkan keputusan yang lebih baik. Kajian ini penting kerana pemaju dalam sektor perumahan dapat menetapkan atau menentukan harga rumah teres dengan lebih stabil dan para pembeli dapat mengetahui harga struktur rumah yang dibina dengan lebih tepat. Masalah ini dapat dilihat dengan kenaikan gelagat harga rumah yang tidak setara dengan pertumbuhan industri perumahan kerana wujud masalah-masalah seperti kepenggunaan tanah, kod dan rancangan bangunan, pengagihan dan pengurusan data kewangan (Razali, 1997) yang terpaksa dihadapi oleh pihak pemaju perumahan.

## METODOLOGI KAJIAN

Terdapat sebanyak sembilan pembolehubah yang digunakan (INSPEN, 1996) untuk menentukan indeks harga rumah iaitu tahun data diambil, jenis tanah dan keluasan kawasan, jenis rumah teres, jenis pegangan, usia rumah, jarak dari bandar, kualiti bangunan dan kualiti kawasan. Data diproses dan ditapis bagi memastikan hanya data yang sesuai digunakan untuk membentuk model. Data yang sesuai bermaksud data yang mempunyai cukup pembolehubahnya, cukup nilainya, berada dalam julat yang relevan dan konsisten dengan harga rumah dalam pangkalan data. Kemudian data yang telah ditapis diwakilkan dan diskalakan supaya diseragamkan dalam satu piawaian untuk memudahkan latihan dan pengujian dilakukan. Seterusnya data tersebut dibahagikan kepada dua bahagian iaitu data untuk latihan dan data untuk pengujian. Data diwakilkan dalam bentuk nombor binari iaitu 0 dan 1. Ini memudahkan rangkaian neural memproses data bagi menghasilkan isyarat output iaitu dalam bentuk magnitud.

Pembolehubah-pembolehubah yang diwakilkan ialah pembolehubah jenis rumah teres, jenis pegangan rumah, usia rumah, kualiti bangunan dan kualiti kawasan. Manakala proses penskalaan adalah menggunakan formula berikut (Skapura, 1995);

$$x_i = \frac{((x_i - \min(x_i)) * (U - 1) + L}{(\max(x_i) - \min)(x_i)}$$

di mana x adalah pembolehubah yang terlibat
U adalah nilai pembolehubah tertinggi
L adalah nilai pembolehubah terendah

Pembolehubah-pembolehubah yang terlibat dalam proses penskalaan iaitu keluasan tanah, keluasan kawasan, jarak dari bandar dan harga rumah. Input data yang digunakan dalam model latihan dibahagikan kepada tiga kumpulan iaitu kumpulan mewakili input latihan, kumpulan data kesahan dan kumpulan data ujian. Dua bentuk model latihan untuk data akan digunakan iaitu;

- Model latihan pertama
  Input data sebanyak 80% digunakan untuk data latihan, 10% untuk data kesahan dan 10% untuk data ujian.

- Model latihan kedua
  Input data sebanyak 70% digunakan untuk data latihan, 20% untuk data kesahan dan 10% untuk data ujian.

Model latihan ini dibuat dalam tiga peringkat iaitu;

a) Peringkat pertama (jujukan)
   Latihan dilakukan dengan input data dipilih secara jujukan satu persatu dan disusun dalam kedudukan menaik.
b) Peringkat kedua (rawak)
   Latihan dilakukan dengan input dipilih secara rawak dengan tidak mengikut susunan.
c) Peringkat ketiga (tapisan)
   Data dipilih secara rawak dan tapisan dibuat ke atas setiap data rawak yang dipilih. Tujuan tapisan adalah untuk memastikan data yang digunakan adalah cukup bagus dalam membentuk model terbaik.

Cara yang terbaik untuk mewakilkan data input dan memilih senibina model rangkaian adalah sewaktu fasa latihan data dijalankan. Ini dilakukan dengan menganalisis proses latihan dan memilih model yang menghasilkan kerangka model yang paling tepat dengan sasaran yang diberi. Kerangka model yang dipilih didapati daripada set-set pemberat yang konsisten setelah semua data dilatih. Kemudian prestasi rangkaian akan diperiksa semasa pengujian sebelum ianya dilaksanakan. Secara prinsipnya, semakin kerap data input digunakan untuk latihan semakin tepat model yang direkabentuk.

Jadual 1 menunjukkan pembahagian secara jelas lokasi data model yang menggunakan tiga jenis fungsian iaitu fungsian linear, fungsian tangen hiperbolik dan fungsian sigmoid. Didapati julat nod yang baik dalam model ini adalah julat di antara nod empat dan nod tujuh dengan menggunakan pendekatan algoritma 'conjugate' sebagai algoritma pembelajaran. Kerangka model mengambilkira tiga pendekatan iaitu pendekatan jujukan, rawak dan tapisan. Daripada latihan data rangkaian secara terperinci, didapati bahawa pendekatan tapisan adalah lebih baik keputusannya berbanding pendekatan secara jujukan dan pendekatan secara rawak. Didapati juga model 22 adalah model terbaik daripada latihan data yang dilakukan dengan penggunaan input data sebanyak 80% untuk latihan, 10% untuk kesahan dan 10% untuk pengujian. Model ini menggunakan fungsian tangen hiperbolik dengan satu lapisan tersembunyi

yang mempunyai lima nod dan menghasilkan 87.08% ketepatan rawak. *'Root Mean Square'* (RMS) yang diperolehi ialah 0.05714 dan ralat min ialah 0.039 iaitu 16.36%.

## HASIL KAJIAN

### Model Latihan

**Jadual 1**
Proses Pemodelan Perceptron Multi Aras

| LOKASI DATA MODEL | STATUS MODEL | | | | RAMALAN (%) | | |
|---|---|---|---|---|---|---|---|
| | MODEL | NOD | FUNGSIAN | ALGO | JUJUKAN | RAWAK | TAPISAN |
| Latihan :70% | 1 | 4 | linear | conj.grad | 67.62 | 73.81 | 77.14 |
| Kesahan :20% | 2 | 5 | linear | conj.grad | 70.00 | 76.67 | 76.86 |
| Pengujian :10% | 3 | 6 | linear | conj.grad | 67.14 | 76.19 | 77.62 |
| | 4 | 7 | linear | conj.grad | 72.38 | 75.71 | 80.00 |
| | 5 | 4 | sigmoid | conj.grad | 78.57 | 86.67 | 86.90 |
| | 6 | 5 | sigmoid | conj.grad | 80.48 | 83.81 | 83.95 |
| | 7 | 6 | sigmoid | conj.grad | 81.43 | 81.43 | 82.52 |
| | 8 | 7 | sigmoid | conj.grad | 76.19 | 79.05 | 80.48 |
| | 9 | 4 | tanh | conj.grad | 72.86 | 83.33 | 83.48 |
| | 10 | 5 | tanh | conj.grad | 70.00 | 83.33 | 84.52 |
| | 11 | 6 | tanh | conj.grad | 76.19 | 80.48 | 82.86 |
| | 12 | 7 | tanh | conj.grad | 70.48 | 79.52 | 80.00 |
| Latihan : 80% | 13 | 4 | linear | conj.grad | 70.42 | 75.42 | 75.83 |
| Kesahan :10% | 14 | 5 | linear | conj.grad | 73.75 | 75.42 | 76.17 |
| Pengujian :10% | 15 | 6 | linear | conj.grad | 60.83 | 75.42 | 75.75 |
| | 16 | 7 | linear | conj.grad | 69.77 | 75.42 | 75.17 |
| | 17 | 4 | sigmoid | conj.grad | 75.00 | 81.25 | 81.42 |
| | 18 | 5 | sigmoid | conj.grad | 78.75 | 83.33 | 84.25 |
| | 19 | 6 | sigmoid | conj.grad | 75.00 | 84.17 | 84.42 |
| | 20 | 7 | sigmoid | conj.grad | 75.42 | 77.08 | 85.83 |
| | 21 | 4 | tanh | conj.grad | 72.92 | 80.50 | 81.00 |
| | 22 | 5 | tanh | conj.grad | 72.08 | 85.00 | 87.08 |
| | 23 | 6 | tanh | conj.grad | 77.92 | 78.75 | 80.00 |
| | 24 | 7 | tanh | conj.grad | 78.33 | 85.83 | 86.25 |

Jadual 1 menunjukkan pembahagian secara jelas lokasi data model yang menggunakan tiga jenis fungsian iaitu fungsian linear, fungsian tangen hiperbolik dan fungsian sigmoid. Didapati julat nod yang baik dalam model ini adalah julat di antara nod empat dan nod tujuh dengan menggunakan pendekatan algoritma 'conjugate' sebagai algoritma

pembelajaran. Kerangka model mengambilkira tiga pendekatan iaitu pendekatan jujukan, rawak dan tapisan. Daripada latihan data rangkaian secara terperinci, didapati bahawa pendekatan tapisan adalah lebih baik keputusannya berbanding pendekatan secara jujukan dan pendekatan secara rawak. Didapati juga model 22 adalah model terbaik daripada latihan data yang dilakukan dengan penggunaan input data sebanyak 80% untuk latihan, 10% untuk kesahan dan 10% untuk pengujian. Model ini menggunakan fungsian tangen hiperbolik dengan satu lapisan tersembunyi yang mempunyai lima nod dan menghasilkan 87.08% ketepatan rawak. *'Root Mean Square'* (RMS) yang diperolehi ialah 0.05714 dan ralat min ialah 0.039 iaitu 16.36%.

## Pengujian Data

**Jadual 2**

Min Ralat Piawai (MSE) Data Ujian

| TAHUN | DATA | MSE | |
|---|---|---|---|
| | | MLP | MR |
| Ke-1 | 100 | 0.0042011 | 0.0046380 |
| Ke-2 | 100 | 0.0017196 | 0.0018250 |
| Ke-3 | 100 | 0.0051654 | 0.0061550 |
| Ke-4 | 100 | 0.0032814 | 0.0035650 |
| Ke-1 | 500 | 0.0025116 | 0.0027920 |
| Ke-2 | 500 | 0.0041026 | 0.0045100 |
| Ke-3 | 500 | 0.0061164 | 0.0064070 |
| Ke-4 | 500 | 0.0072126 | 0.0075370 |
| Ke-1 | 600 | 0.0089205 | 0.0093420 |
| Ke-2 | 600 | 0.0039433 | 0.0040280 |
| Ke-3 | 600 | 0.0068772 | 0.0070710 |
| Ke-4 | 600 | 0.0081225 | 0.0085740 |

Set pengujian data dilakukan untuk melatih rangkaian dalam pengujian perlaksanaan ramalan bagi mendapatkan model yang terbaik. Dalam Jadual 2, data diuji dalam tiga jumlah kategori iaitu 100 data, 500 data dan 600 data. Data bagi setiap kategori untuk setiap tahun diuji dengan menggunakan pendekatan rangkaian neural jenis perceptron multi aras (MLP) dan regresi berganda (MR). Persamaan untuk model regresi berganda yang didapati daripada input data adalah seperti berikut;

$$y = 0.0317x_1 + 0.0183x_2 + 0.3142x_3 + 0.0595x_4 + 0.5266x_5 + 0.0332x_6 - 0.974x_7 + 0.0263x_8 - 0.09549x_9 - 0.293$$

dengan pembolehubah bersandar y mewakili harga rumah. Manakala pembolehubah tidak bersandar adalah tahun, keluasan tanah, jenis rumah teres, jenis pegangan tanah, keluasan bangunan, usia rumah, jarak rumah dari bandar, kualiti kawasan dan kualiti bangunan yang diwakili oleh $x_i$,i = 1, 2, 3, ..., 9. Semua nilai min ralat piawai (MSE) daripada dua pendekatan ini didapati kurang daripada 0.01. Didapati bahawa pendekatan perceptron multi aras menghasilkan keputusan yang lebih baik berbanding dengan pendekatan regresi.

## KESIMPULAN

Dalam kajian ini, penekanan tentang aktiviti tapisan ke atas data semasa tahap prapemprosesan dalam pendekatan rangkaian neural berasaskan perceptron multi aras telah dilaksanakan. Aktiviti tapisan yang telah dipraktikkan pada tahap prapemprosesan data telah menyumbang kepada pembentukan model yang lebih baik jika dibandingkan dengan model yang dibentuk daripada data yang tidak dilaksanakan aktiviti tapisan. Seterusnya keputusan yang diperoleh daripada model rangkaian neural telah menunjukkan prestasi ramalan yang lebih baik jika dibandingkan dengan keputusan daripada analisis regresi berganda. Secara kesimpulannya, dengan memasukkan unsur teknik tapisan dalam prapemprosesan, rangkaian neural akan dan boleh menghasilkan keputusan ramalan dengan lebih tepat berbanding analisis regresi berganda.

## RUJUKAN


Bahrom, S., & Shang, S.W. (1996). Neural network approach in predicting KLSE composite index. *Laporan Teknik Jabatan Matematik,* Universiti Teknologi Malaysia.

Bode, J. (1998). Neural networks for cost estimation. *Cost Engineering*, Jan, 40(1), 25 – 30.

Cybenko, G. (1989). Approximation by superposition of a sigmoidal function. *Math. Control Signals Systems*, 2, 303.

Famili, A., Wei, M.S., Richard, W., & Evangelos, S. (1997). Data preprocessing and intelligent data analysis. *Intelligent Data Analysis*, 1(1).

Herbert, W. (1998). Economic prediction using neural networks: The case of IBM daily stock returns. *Proceeding of the IEEE International Conference on Neural Networks* : II1451-II450.

Hartman, E., & Keeler, J. D. (1991). Predicting the future: advantages of semilocal units. *Neural Computing*, 3, 566.

INSPEN (1996). *Malaysia House Price Index : Technical Summary*. Kuala Lumpur: Kementerian Kewangan Malaysia

Ivo, D., & Gunter, G. (1998). Simple data filtering in rough set systems. *International Journal of Approximate Reasoning*, 18, 93-106.

Ku-Mahamud, K.R. , Abu Bakar, A., & Nawawi, N. (1999). Multi layer perceptron modelling in housing market. *Malaysian Management Journal*, 3(1), 61-69.

Lönnblad, L. , Peterson, C., & Rögnvaldsson, T. (1992). Mass reconstruction with a neural network. *Phys. Lett.* , B278, 181.

Marcus, D.O., & Ramesh, S. (1990). A neural network model for bankruptcy prediction. *Proceeding of the IEEE International Conference on Neural Networks,* San Diego, Jun *; II163-II168.*

Law, R. (1998). Room occupancy rate forecasting: a neural network approach. *International Journal of Contemporary Hospitality Management*, 10(6), 234 – 239.

Lou, M. (1993). Preprosessing data for neural network, *Technical Analysis of Stock & Commodities*, © 1993 Technical Analysis, Inc.

Peterson, C. , Rögnvaldsson, T., & Lönnblad, L. (1994). JETNET 3.0 - a versatile artificial neural network. *Package Computer Physics Communications*, 81, 185-220.

Ohlsson, M. , Peterson, C. , Pi, H. , Rögnvaldsson, T. & Söderberg, B. (1994). Predicting Utility Loads with Artificial Neural Networks — Methods and Results from the Great Energy Predictor Shootout, Lund Preprint LU TP 93-24.

Razali, A. (1997). *Housing the Nation : A Definitive Study*. Kuala Lumpur: *Cagamas Berhad.*

Skapura, D.M. (1995*). Building Neural Networks*. New York: Addison Wesley.

Soumitra, D., & Sashi, S. (1998). Bond rating: A non-conservative application of neural network. *Proceeding of the IEEE International Conference on Neural Networks,* II1443-II450.

Wang, H., & Ho, K.H. (1995). Artificial intelligent modelling of the private housing market in Singapore. *Proceeding of the International Congress on Real Estate*, Singapore, April.