國立成功大學
工業與資訊管理學研究所

類神經期末報告
報告論文：

以逐步分解迴歸分析法建構房地產估價模型

Building real estate valuation models with comparative approach through case-based reasoning

I-Cheng Yeh, Tzu-Kuang Hsu

組別：一

組員：

徐致恆
徐心縈
張雅婷
陳昱嘉

目錄

一、 簡介

1. 報告摘要：

為了解探討如何有效地應用類神經網路於房地產估價上，論文利用改善傳統多變數迴歸分析法的逐步分解迴歸分析，來針對新店區 2012 年 6 月~2013 年 5 月之房地產交易資料進行測試，而本篇報告本組將比較 LSTM 及 GRU 與逐步分解迴歸分析法三種模式於估價精確度上之差異，並使用 RMSE 和 R Squared 做為衡量模型效能，以為未來改進房地產估價輔助系統之參考。

2. 本論文方法:

本論文將原本傳統的多變數迴歸分析法

$$Y = a_0 + \sum_{i=0}^{m} a_i X_i$$

改為逐步分解迴歸分析

$$Y_j = \prod_{i=1}^{m} k_{ij} \times \overline{Y}$$

增加了易理解性和改善了能夠表達自變數與因變數之間非線性關係的能力的優點。

3. 論文數據:

2012 年 6 月至 2013 年 5 月年新店區房屋買賣數據，包含: 交屋年月、屋齡、距離最近捷運站的距離、徒步生活圈內的超商數、緯度、經度六個屬性來預測房價類別值。

二、 文獻模型

1. 模型步驟流程介紹

此文獻提出一個逐步分解迴歸分析的新研究方法，假設房地產每坪的單價是供需圈內每坪單價與多個屬性的無因次的調整係數的連乘積，從最重要的屬性開始，逐一分解各個屬性的調整係數估計值，並建構各調整係數的預測模型，步驟如下：

(1) 排序：以排序等分法估計屬性的重要性。

(2) 分解：以逐步分解法建構各屬性的調整係數之單變數迴歸模型。

(3) 整合：將各屬性之調整係數迴歸模型整合為價格預測模型。

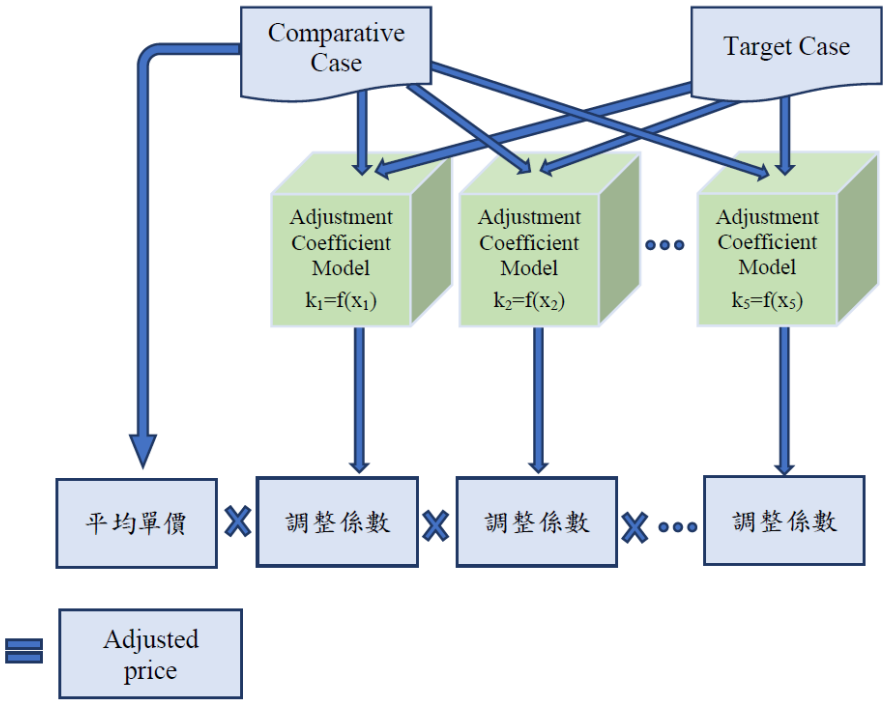


圖 1、 將各屬性的調整係數之迴歸模型整合為價格預測模型

排序等分法：

(1) 將樣本依照屬性排序，並分成十等分。

(2) 計算每一個等分的樣本的目標變數平均值。

(3) 計算不同等分的目標變數平均值的標準差。

目標變數依照屬性排序等分後，計算目標變數平均值的標準差，表該屬性越能解釋目標變數，即該屬性重要性越高。由表 1 可知新店區變異較大的屬性，故最佳預測模型為：

(1) 距離最近捷運站距離 k1：對數模型

(2) 徒步可到之超商數 k2：線性模型

(3) 屋齡 k3：對數模型

(4) 交易日期 k4：線性模型

表 1、各參數變異之均方根

	距離最近捷運站距離	徒步可到之超商數	屋齡	交易日期
RMSE	10.6	8.3	6.8	2.9

表 2、各屬性之調整係數模型

	係數 a	係數 b	判定係數	誤差均方根
k1	-0.23846	2.527609	0.8820	0.081076
k2	0.009452	0.955591	0.2830	0.018765
k3	-0.06408	1.159766	0.5238	0.036351
k4	0.21170	-20.62480	0.7283	0.013932

結論：新店區逐步分解迴歸法模型判定係數為 0.6636，誤差均方根為 7.95(萬元/坪)，平均房價為 37.98(萬元/坪)。

2. 小結：

逐步分解迴歸研究限制:

(1) 房地產類型只限於公寓與大樓的住宅商品。

(2) 研究地域只限於台北市與新北市有捷運可到的都會區。

(3) 由於此研究資料取得的限制，根據資料屬性重要性取五個因子來建構模型。

(4) 此研究之資料只限於 2012 年 6 月至 2013 年 5 月。

(5) 基於保護隱私權，實價登錄資料庫內數據的門牌號碼並非完全精確的地址，只顯示了門牌號 50 號以內的範圍，門牌範圍並無規定的大小，有可能影響座標空間因子的精度。

三、 深度學習應用

住房價格會因為住房本身條件不同，受到外在環境及時間因子持續發生變動，因此應用深度學習的循環類神經網路(RNN)的長短期記憶(LSTM)演算法和 GRU 去建構住房價格估價模型。接著將針對這三種模型進行進一步介紹。

1. 循環類神經網路(Recurrent Neural Network, RNN)

循環類神經網路會將前一時間的資訊帶到下一時間，除了讓神經網路有類似記憶性功能之外，也適合辨別時間前後有關連性的資料，本論文中房地產的成交年月就是此性質資料。

下面我們可以看到，網絡將前一時間步的網絡輸出作為輸入，並使用前一時間的內部狀態作為當前時間的起點。

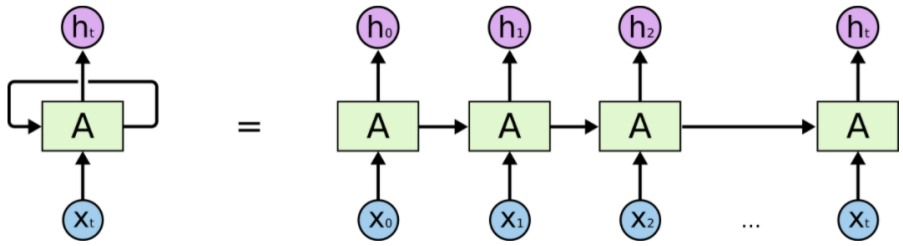


圖 2、循環類神經網路架構圖

循環類神經網路因每個時間都有輸入和輸出，因此可依照任務不同，決定要如何選取輸入資料以及輸出資料(結果)。

縱使循環類神經網路藉著循環的結構使模型擁有類似人的記憶性質，但他存在著無法避免的缺陷，那就是梯度消失和梯度爆炸。當一個循環神經網路倒傳遞修正參數時，且梯度為一個小於 1 的數，當時間步很多時，會因為多次相乘而收斂，此現象稱為梯度消失(gradient vanishing); 反之，當一個循環神經網路倒傳遞修正參數時，且梯度為一個大於 1 的樹，同樣的當 time step 時，會因為多次相乘而發散，此現象稱為梯度爆炸(gradient explored)。

梯度消失與梯度爆炸，直接導致距離當前時間步越遠的記憶被過度的減少，導致倒傳遞時無法調整參數，進而使訓練被迫中斷，為了解決此問題，Long Short-Term Memory 模型的誕生。

2. 長短期記憶(Long Short-Term Memory, LSTM)

LSTM(Long short-term memory)，論文首次發表於 1997 年。由於獨特的設計結構，LSTM 適合於處理和預測時間序列中間隔和延遲非常長的重要事件。主要改善了以前 RNN 的一些問題，首先新增貫穿全局的狀態方程式c_t，這降低了整體模型記憶對h_t的依賴性。

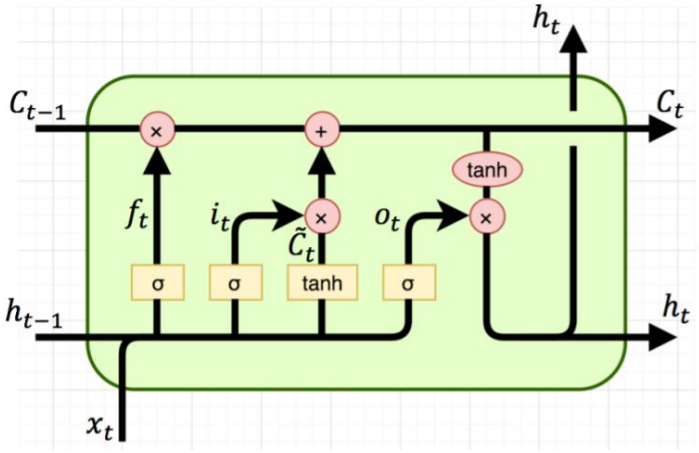


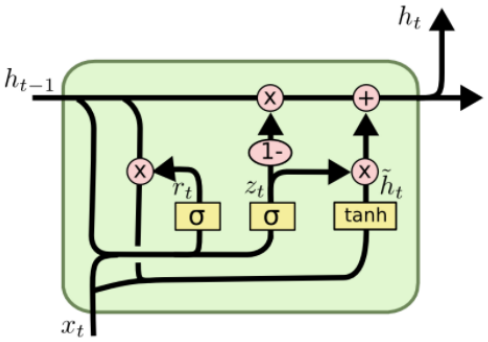
圖 3、 LSTM 架構圖

而 LSTM 由四個 unit 組成: Input Gate、Output Gate、Memory Cell 以及 Forget Gate。進而大幅提高了其在長期記憶的表現。

Input Gate 為當資料輸入時，input gate 可以控制是否將這次的值輸入，並運算數值。Memory Cell 為將運算出的數值記憶起來，以利下個 cell 運用。Output Gate 為控制是否將這次計算出來的值 output，若無此次輸出則為 0。Forget Gate 為控制是否將 Memory 清掉(format)。

3. GRU (Gated Recurrent Unit)

GRU(Gated Recurrent Unit)為 LSTM 的改良模型，去解決 LSTM 執行速度過慢的問題，於 Junyoung,Chung(2014)在"Empirical evaluation of gated recurrent neural networks on sequence modeling."所提到的，GRU 可用來加快執行速度及減少記憶體的耗用



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

圖 4、 GRU 架構圖

因此，為了解探討如何有效地應用類神經網路於房地產估價上，本研究設計實驗，比較 LSTM 及 GRU 與逐步分解迴歸分析法三種模式於估價精確度上之差異，並使用 RMSE 和 R Squared 做為衡量模型效能，以為未來改進房地產估價輔助系統之參考。

四、 實驗分析

1. 衡量指標

(1) R-square(R 平方係數)：

R-square 在本研究用來衡量房地產推估價格與實際成交價格之間關係程度的指標。當 R-square 值越接近 1 時，代表房地產推估價格與實際成交價格可解釋程度越高，反之，當 R-square 值越接近 0 時，代表房地產推估價格與實際成交價格可解釋程度越低。

(2) RMSE(均方根誤差)：

可以用來衡量預測值和實際值之間的偏差，藉此估計預測模型預測目標值的準確度。當 RMSE 越趨近於 0，表示模型之預測能力越佳。

(3) MAPE(平均絕對誤差百分比) ：

可以用來衡量一個模型預測結果的好壞。表示預測值和實際值之間的平均絕對差百分比，當 MAPE 越低，表示模型之預測能力越佳。

2. 資料處理

(1) UCI 資料庫

本研究使用 UCI 資料集中的 Real estate valuation data set Data Set 的資料，時間為 2012 年 6 月至 2013 年 5 月，以新店區的公寓與大樓住宅作分析，樣本數共有 414 筆，其中依照交易日期排序後，取資料庫前 80%筆數作為訓練資料(Training data)共 331 筆；資料庫後 20%作為測試資料(Testing data)共 83 筆。

(2) 選擇變數說明

變數選擇為交屋年月、屋齡、距離最近捷運站的距離、徒步生活圈內的超商數、緯度、經度等六個屬性。以下表 3 是此研究之原始資料。

表 3、原始資料

No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude
1	2012.917	32	84.87882	10	24.98298	121.54024
2	2012.917	19.5	306.5947	9	24.98034	121.53951
3	2013.583	13.3	561.9845	5	24.98746	121.54391
4	2013.500	13.3	561.9845	5	24.98746	121.54391
5	2012.833	5	390.5684	5	24.97937	121.54245
6	2012.667	7.1	2175.03	3	24.96305	121.51254
7	2012.667	34.5	623.4731	7	24.97933	121.53642
8	2013.417	20.3	287.6025	6	24.98042	121.54228
9	2013.500	31.7	5512.038	1	24.95095	121.48458
10	2013.417	17.9	1783.18	3	24.96731	121.51486
11	2013.083	34.8	405.2134	1	24.97349	121.53372
12	2013.333	6.3	90.45606	9	24.97433	121.5431
13	2012.917	13	492.2313	5	24.96515	121.53737
14	2012.667	20.4	2469.645	4	24.96108	121.51046
15	2013.500	13.2	1164.838	4	24.99156	121.53406
16	2013.583	35.7	579.2083	2	24.9824	121.54619

接著將數值資料做正規化，並利用交易日期進行排序，呈現為表 4。

表 4、原始資料正規化後數值表

1	X1 transaction date'	'X2 house age'	'X3 distance to the nearest MRT	'X4 number of convenience stores'	'X5 latitude'	'X6 longitude'
2	0	0.1621	0.332833	0.3	0.375424	0.420638
3	0	0.787671	0.092827	0.7	0.57271	0.678132
4	0	0.465753	0.378407	0.4	0.351551	0.39821
5	0	0.034247	0	0.7	0.432016	0.727733
6	0	0.070776	0.055762	0.5	0.591129	0.758896
7	0	0.070776	0.085786	0.6	0.484004	0.794587
8	0	0.671233	0.69408	0.1	0.208192	0.236036
9	0	0.746575	0.05709	0.6	0.386694	0.743692
10	0	0.69863	0.018633	0.8	0.599612	0.732047
11	0	0.762557	0.025305	0.6	0.411658	0.739487
12	0	0.294521	0.072525	0.5	0.400873	0.688376
13	0	0.796804	0.024142	0.8	0.501939	0.743153
14	0	0.353881	0.122598	0.4	0.567014	0.658939

3. 模型建構

圖 5 為 LSTM 的訓練收斂狀況，如圖所示，訓練誤差 RMSE 為 0.076316，epoch 大約 20 次即可將 LSTM 收斂。

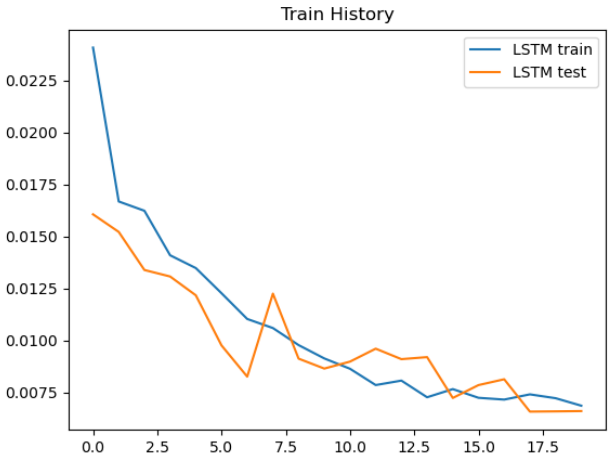


圖 5、LSTM 學習資料收斂圖

圖 6 為 GRU 的訓練收斂狀況，如圖所示，訓練誤差 RMSE 為 0.077222，epoch 大約 20 次即可將 GRU 收斂。

圖 6、GRU 學習資料收斂圖

表 5 為 LSTM、GRU 與迴歸預測結果之比較，利用衡量指標，來比較不同方法之間的效能。

表 5、類神經網路與迴歸預測結果比較

衡量指標 / 模型	MAPE%	RMSE	R-Squared	預測平均	RMSE/平均
逐步分解迴歸	無	7.73	68.3%	37.98	0.204
迴歸	28.76%	0.08598	51.78%	0.26749	0.32144
GRU	23.64%	0.07632	62.01%	0.27252	0.28004
LSTM	23.39%	0.07722	61.11%	0.26319	0.2934

經由比較結果發現，本次研究使用的 GRU 預測 MAPE=23.64%，LSTM 預測 MAPE=23.39%，而迴歸預測 MAPE=28.76%。而 R-Squared 方面，GRU 預測 R-Squared=62.01%，LSTM 預測 R-Squared=61.11%，而迴歸預測 R-Squared=51.78%，由此可知類神經預測能力稍比迴歸預測結果佳。

五、 結論與建議

1. 結論

(1) 逐步迴歸的表現最佳

逐步迴歸由於對每個屬性都做模型配適，可對每個屬性作出最好的解釋，例如距離最近捷運站距離 k1：對數模型、徒步可到之超商數 k2：線性模型、屋齡 k3：對數模型交易日期 k4：線性模型、並針對經緯度有結合兩個屬性作出一個空間因子，做出解釋，再將每個係數相乘。

而多元迴歸每個自變數跟應變數只有線性模型，而無法精準匹配，所以得出逐步分解迴歸的表現較佳。

(2) 迴歸模型和類神經模型之間的比較

由於影響房價的因子論文中有十幾個，論文中作者挑了其中五個去進行預測，這五個很可能是作者已經事前挑選過可以讓逐步分解表現得最好的五個屬性。

而我們類神經模型選擇作者給的資料集，所以很可能有對我們類神經更好的屬性沒挑到，例如沒有經過作者的空間因子轉換，本組的類神經直接接收 raw data 之經度、緯度資料是對房價沒有甚麼顯著的解釋能力的，只能靠前面幾個屬性解釋。

(3) 迴歸具有良好的解釋能力，不管是逐步迴歸的相乘係數還是多元迴歸的相加係數，都很好的解釋當屬性增加或減少多少時會對應變數房價造成多少改變，而類神經就缺乏可解釋性。

(4) LSTM 和 GRU 在本的表現在本資料集中表現差異不大

由於本資料集之資料量不大，共有 414 筆，所以在時間或學習效果的表現差異都不大。

2. 建議

(1) 本次研究因為時間和資料因素的限制，時間只限 2012 年 6 月至 2013 年 5 月，資料筆數只有 414 筆，若要提高模型的準確率，勢必要增加研究的時間長度，並增加資料量，來修正模型。

(2) 本次研究的五個屬性，是原論文作者所選擇對房價有影響力之因子，若要做進一步研究，可增加其他有影響力的因子，其他如每坪公告地價、距離大賣場的距離、徒步生活區是否有校園等因子，此外，本研究屬性中的徒步生活圈的超商數，只針對統一超商，未來可考慮其他超商。

(3) 本次研究指針對新店區公寓與大樓的住宅商品進行分析，若進一步探討，如果以透天厝做為研究，類神經網路是否具備更高的預測能力。

(4) 類神經網路雖然有優秀學習能力，但在選擇隱藏層層數、隱藏層單位數目、轉換函數、迭代次數等變數影響，需要花時間去確認，另外，樣本數如果不足，也可能出現擬合過度(overfitting)問題，而失去預測準確性。

(5) 本次研究的每個屬性權重相同，亦指沒有考慮個屬性的權重關係，若之後要進行進一步研究，可以去分析各個屬性對房價的影響程度，並進行權重分析。

(6) 房地產價格可能會受到面臨道路寬度、鄰近有無嫌惡設施(焚化廠、變電所、工廠、高壓電塔)、鄰近有無友善設施等影響，此影響可能造成類神經網路預測與實際情況有落差，因此之後進一步研究可將相關信息加以考慮，以減少預測誤差。

(7) 本研究使用之類神經模型為 LSTM 和 GRU，因本次研究的樣本數及屬性較少，且時間序列影響不大，可以採用其他模型進行預測，進而去探討不同的模型是否有較好的預測結果。

參考文獻

一、中文

1.謝孟勳(2017)，實價登錄資料庫結合類神經網路推估房地產市價，國立中興大學土木工程學系碩士論文

2. 葉怡成, 丁導民, & 詹巧薇. (2016)，以逐步分解迴歸分析法建構房地產估價模型. 營建管理季刊, (105), 54-70.

3.蔡瑞煌, 高明志, & 張金鶚. (1999). 類神經網路應用於房地產估價之研究. 住宅學報, (8), 1-20.