國立雲林科技大學資訊管理系

機器學習

專案作業一

USING FEED-FORWARD NEURAL NETWORK TO PREDICT ONLINE SHILL BIDDING BEHAVIOR AND EDUCATION LEVEL OF ADULTS

以前饋式神經網路預測網路競標行為以及成人之教育程度

M10723047 黃靖媛

M10923012 陳羿欣

M10923021 吳青芬

M10923023 張凱淇

指導教授：許中川

2021 年 04 月

摘要

在本研究使用 Shill Bidding DataSet 與 Adult Dataset，作為實驗資料集，在 Shill Bidding DataSet 進行兩項預測任務，分別為二分類預測以及數值預測，任務一主要希望透過神經網路之訓練模型找出不正常拍賣行為，任務二主要為預測拍賣者之得標率，在 Adult Dataset 中為收集之成年人口普查收入之相關數據來進行多分類預測任務，以此來預測成年人的教育程度。在本研究中分類預測模型利用 Precision、Recall 及 F1-score；數值預測模型使用 MAE、MAPE 與 RMSE 來評估模型，經由評估指標計算顯示，本研究設置之神經網路模型在 Shill Bidding DataSet 任務能符合預測任務之需求，且發現模型設置之節點數與層數，會影響預測模型之績效，而在 Adult Dataset 則無法獲得較好的績效。

關鍵字：類神經網路、網路競標、教育程度、預測

一、緒論

1.1 動機

在行銷領域中，企業常利用蒐集得來的客戶消費相關數據來得知該客戶之消費行為，針對消費者習慣作不同的行銷策略，或是異常的消費行為，如購物詐欺等，現今網路商機無窮，購物詐欺行為常出現於網路競標之中，因此本研究使用 Shill Bidding Dataset 公開資料集，該資料集為關於在 eBay 上拍賣了大量流行產品，這個資料集提供者收集了關於拍賣行為中的各種動作和紀錄。在拍賣上惡意賺錢者可能會犯三種欺詐行為，即拍賣前欺詐，例如黑市商品拍賣，競標期間發生的拍賣中欺詐，例如 Shill Bidding（SB）以及拍賣後欺詐（Alzahrani & Sadaoui，2020）。因此本研究希望透過 eBay 的交易紀錄來預測交易行為屬於正常交易抑或是非正常交易，進而阻止詐欺交易的發生，本研究利用神經網路預測拍賣者之行為是否為正常的競標以及該拍賣者之得標率。

在第二個資料集成人資料集（Adult Dataset）所收集之人口普查收入數據，現今社會幾乎人人都有一張大學文憑，在某篇新聞中報導一位博士生畢業後在賣雞排，而有位企業家教育程度不高，但如今他成為了成功的企業家，因此本研究想透過成人資料集（Adult DataSet）來預測教育程度，藉此來了解教育程度的高低是否會受其他因素影響。

1.2 目的

在本研究使用 Shill Bidding DataSet 與 Adult Dataset，作為實驗資料集，在 Shill Bidding DataSet 進行兩項預測任務，分別為二分類預測以及數值預測，任務一主要希望透過神經網路之訓練模型來預測拍賣行為是否為正常交易，利用 Precision、Recall 及 F1 指標來評估模型是否能找出可能是犯罪類的資料，以此得知模型是否能在新的資料預測出是否是詐欺交易，來預防欺詐拍賣的發生；任務二主要為預測拍賣者之得標率，使用 MAE、MAPE 與 RMSE 來評估模型。

在 Adult Dataset 中為收集之成年人口普查收入之相關數據，進行教育程度之預測任務，利用 Precision， Recall 及 F1 指標評估，以此來預測成年人的教育程度。

二、方法

2.1 實作方法說明

本研究實驗資料集包括 Shill Bidding DataSet、Adult Dataset，首先將資料進行前置處理，其包括資料清理、Ono-hot encoding、資料切割，即為將資料分成訓練資料（train data）以及測試資料（test data）。

Adult 資料集中包含“?”、空白值等的資料，因此在進行資料分割前先清理資料雜訊，在此資料集主要預測成人之教育程度，即為在資料集中“education”欄位，而另一個資料集 Shill Bidding Data Set 要去預測是否正常交易以及得標率，即為“Class”、“Winning_Ratio”欄位，兩資料集皆使用前饋式類神經網路來進行類別與數值的預測，再對類別類型資料使用 Precision、Recall 及 F1 等指標對模型進行優劣評估，數值類則使用 MAE、MAPE 與 RMSE 等指標對模型進行優劣評估。

2.2 程式執行方法說明

本研究利用 Anaconda3 的 Jupyter notebook 環境來進行開發，使用 pandas 和 numpy 套件對 Adult Data Set 和 Shill Bidding Data Set 進行資料的前處理，最後將資料分割放到使用 keras 套件建立前饋式類神經網路模型，對資料進行預測，再將模型建立時所記錄的每個 epoch 訓練後的 loss 和 accuracy 畫出檢視學習歷程是否正常後再對類別類型資料使用 Precision，Recall 及 F1 等指標對模型進行優劣評估，數值類則使用 MAE、MAPE 與 RMSE 等指標對模型進行優劣評估。

三、實驗

3.1 資料集

本研究使用 Shill Bidding Dataset 以及 Adult Dataset 兩種資料集作為實驗資料，以下為兩種資料集前處理之前之規格與說明。

3.1.1 Shill Bidding 資料集

Shill Bidding 資料集為在 eBay 拍賣會上之相關數據，如競標者 ID、競標者趨勢、競標比率、拍賣持續時間、獲勝率等等，包含了數值資料及非數值資料。

- 原始資料筆數：6321
- 正規化後之訓練資料筆數：5056
- 正規化後之測試資料筆數：1265

表 1

Shill Bidding 資料集 部分原始資料

欄位名稱 / 資料編號	Record_ ID	Auction_ ID	Bidder_ ID	Bidder_ Tendency	Bidding_ Ratio	…	Winning_ Ratio	Auction_ Duration	Class
410	1	732	_***i	0.2	0.4	…	0.666667	5	0
1	2	732	g***r	0.02439	0.2	…	0.944444	5	0
2	3	732	t***p	0.142857	0.2	…	1	5	0
3	4	732	7***n	0.1	0.2	…	1	5	0
4	5	900	z***z	0.051282	0.222222	…	0.5	7	0
5	8	900	i***e	0.038462	0.111111	…	0.8	7	0
6	10	900	m***p	0.4	0.222222	…	0.75	7	0
7	12	900	k***a	0.137931	0.444444	…	1	7	1
8	13	2370	g***r	0.121951	0.185185	…	0.944444	7	1
9	27	600	e***t	0.155172	0.346154	…	0.611111	7	1

3.1.2 Adult 資料集

Adult 資料集為 Barry Becker 從 1994 年人口普查數據庫中收集而成，該資料集的主要任務為分類成人年薪資是否為 5 萬元，但在本研究之實驗，將之應用於類神經網路學習上，並且預測成人之教育程度。

- 原始資料筆數：48842
- 正規化後之訓練資料筆數：30162

- 正規化後之測試資料筆數：15059

表 2

Adult 資料集欄位介紹

欄位	屬性	內容
0	age	continuous
1	workplace	Private，Self-emp-not-inc，Self-emp-inc，Federal-gov，Local-gov，State-gov，Without-pay，Never-worked
2	fnlwt	continuous
3	education	Bachelors，Some-college，11th， HS-grad，Prof-school，Assoc-acdm，Assoc-voc，9th，7th-8th，12th，Masters，1st-4th，10th，Doctorate，5th-6th，Preschool
4	education-num	continuous
5	marital-status	Married-civ-spouse，Divorced，Never-married，Separated，Widowed，Married-spouse-absent，Married-AF-spouse
6	occupation	Tech-support，Craft-repair，Other-service，Sales，Exec-managerial，Prof-specialty，Handlers-cleaners，Machine-op-inspct，Adm-clerical，Farming-fishing，Transport-moving，Priv-house-serv，Protective-serv，Armed-Forces
7	relationship	Wife，Own-child，Husband，Not-in-family，Other-relative，Unmarried
8	race	White，Asian-Pac-Islander，Amer-Indian-Eski-moOther，Black
9	sex	Female，Male
10	capital-gain	continuous
11	capital-loss	continuous
12	hours-per-week	continuous
13	native-country	United-States Cambodia，England，Puerto-Rico，Canada，Germany，Outlying-US(Guam-USVI-etc)，India，Japan，Greece，South，China，Cuba，Iran，Honduras，Philippines，Italy，Poland，Jamaica，Vietnam，Mexico，Portugal，Ireland，France，Dominican-Republic，Laos，Ecuador，Taiwan，Haiti，Columbia，Hungary，Guatemala，Nicaragua，Scotland，Thailand，Yugoslavia，El-Salvador，Trinadad&Tobago，Peru，Hong， Holand-Netherlands
14	salary	<=50K，>50K

表 3

Adult 資料集 部分原始資料（欄位編號對照錯誤! 找不到參照來源。）

欄位編號 / 資料編號	0	1	2	3	4	…	10	11	12	13	14
0	39	State-gov	77516	Bachelors	13	…	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	…	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	…	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	…	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	…	0	0	40	Cuba	<=50K

3.2 前置處理

3.2.1 Shill Bidding 資料集

圖 1 Shill Bidding 資料集-前置處理流程圖

- 資料清理：將資料中與預測結果不相關資訊過濾。
- 資料分割：將資料之類別與欲用來訓練之資料分割，以 8:2 比例分割成 5056 筆訓練資料以及 1265 筆測試資料。

表 4

Shill Bidding 資料集-資料前處理後部分資料

欄位名稱 / 資料編號	Bidder_ Tendency	Bidding_ Ratio	Successive_ Outbidding	Last_ Bidding	Auction_ Bids	…
1	0.200000	0.400000	0.0	0.000028	0.0	…
2	0.024390	0.200000	0.0	0.013123	0.0	…
3	0.142857	0.200000	0.0	0.003042	0.0	…
4	0.100000	0.200000	0.0	0.097477	0.0	…
5	0.051282	0.222222	0.0	0.001318	0.0	…

3.2.2 Adult 資料集

圖 2 Adult 資料集-前置處理流程圖

- 資料分割：將資料之類別與欲用來訓練之資料分割，以 8:2 比例分割成 30162 筆訓練資料以及 15060 筆測試資料。
- 資料清理：將資料中的缺失值以及過濾與預測結果不相關的資訊。
- One-hot encoding：將非數值的屬性進行特徵數值化，。

表 5

Adult 資料集資料前處理後部分資料

欄位名稱 / 資料編號	age	hours_per_week	workclass_Federal-gov	...	sex_Female	sex_Male	high_income_<=50K	high_income_>50K
0	39	40	0	...	0	1	1	0
1	50	13	0	...	0	1	1	0
2	38	40	0	...	0	1	1	0
3	53	40	0	...	0	1	1	0
4	28	40	0	...	1	0	1	0

3.3 實驗設計

以下為 Shill Bidding Dataset 與 Adult Dataset 兩資料集依序的流程步驟：

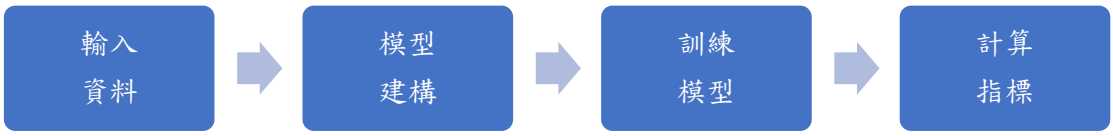

圖 3 實驗流程圖

3.3.1 Shill Bidding Dataset

在 Shill Bidding Dataset 中進行兩項預測任務，一為預測拍賣者是否為正常競標行為，屬於二分類任務；二為預測拍賣之得標率，屬於數值預測任務。兩項任務在模型設置皆使用 Dropout，參數設置為 0.5，使模型在訓練時去除不必要之特徵以及防止模型 overfitting； Batch size 之設置會影響模型優化程度以及收斂速度，由於本資料集之訓練資料筆數為 5056 筆，因此 Batch size 參數設置

為 128，將訓練資料每一回合抽樣 40 次，共抽樣 1000 回合（epochs），若模型訓練在 10 回合 Loss 尚未降低，則提前結束訓練（early stopping），在本實驗中模型之 Optimizer function 原為 RMSprop，在實驗過程中發現在訓練模型時其 Loss 降低過程極為不穩定，因此將其改為廣泛使用之 Adam，經實驗證實其訓練過程穩定許多。

任務一：預測是否為正常競標行為

本實驗資料集資料筆數較少，因此在 Input layer 設定為 64 個 nodes，隱藏層節點設置數量隨著二階層遞減，若將其節點數設為較多如 128 個 nodes，其準確度會較 64 個 nodes 來得低，增加模型層數也是如此，因此在本文件僅顯示模型績效較優之實驗結果。在模型之 Activation function 為 Sigmoid，其餘層為使用較常使用之 Relu 函數，神經網路架構如表 6。

1. 模型建構：設定模型初始參數
- Epochs：1000
- Batch Size：128
- Activation: Sigmoid
- Loss：Binary cross entropy
- Optimizer：Adam
- Metrics：Accuracy, MAE

2. 訓練模型：將資料匯入模型中並進行訓練
3. 計算指標：使用 Precision, Recall 及 F1 指標做為模型評估指標
4. 輸出預測績效結果。

表 6

Shill Bidding Dataset 任務一分類預測模型架構表

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 64)	576
dense_2 (Dense)	(None, 32)	2080
dense_3 (Dense)	(None, 32)	1056
dropout_1 (Dropout)	(None, 32)	0
dense_4 (Dense)	(None, 16)	528
dense_5 (Dense)	(None, 1)	17
Total params: 4.257		
Total params: 4,257		
Non-trainable params: 0		

任務二：預測拍賣之得標率

本任務主要為數值預測，Input layer 與隱藏層節點數與任務一相同，由於本任務主要為預測拍賣得標率（數值預測），因此在模型之 Activation function 將不

設置，其餘層為使用較常使用之 Relu 函數。

1. 模型建構：設定模型初始參數

- Epochs：1000
- Batch Size：128
- Activation : 不設置
- Loss：MSE
- Optimizer：Adam
- Metrics：MAE

2. 將資料匯入模型中並進行訓練
3. 使用 MSE, RMSE, MAE 指標做數值預測
4. 輸出預測績效結果

表 7

Shill Bidding Dataset 任務二數值預測模型架構表

Layer (type)	Output Shape	Param #
dense_6 (Dense)	(None, 64)	512
dense_7 (Dense)	(None, 32)	2080
dense_8 (Dense)	(None, 32)	1056
dropout_2 (Dropout)	(None, 32)	0
dense_9 (Dense)	(None, 16)	528
dense_10 (Dense)	(None, 1)	17
Total params: 4,193		
Total params: 4,193		
Non-trainable params: 0		

3.3.2 Adult Dataset

在 Adult Dataset 中進行教育程度之預測模型，其教育程度之級別共有 16 個級別，屬於多分類任務。在模型設置 Dropout，參數設置為 0.5，Batch size 之設置會影響模型優化程度以及收斂速度，由於本資料集之訓練資料筆數為 30162 筆，因此 Batch size 參數設置為 128，將訓練資料每一回合抽樣 236 次，共抽樣 1000 回合（epochs），若模型訓練在 10 回合 Loss 尚未降低，則提前結束訓練（early stopping），在本實驗中模型之 Optimizer function 廣泛使用之 Adam，在模型之 Input layer 設定為 64 個 nodes，隱藏層節點設置數量隨著二階層遞減，而在模型之 Activation function 為 Sigmoid，其餘層為使用較常使用之 Relu 函數，最後輸出層使用 softmax，其神經網路架構如表 8。

1. 設定模型初始

- Epochs：1000
- Batch size：128

- Activation : Sigmoid, Relu, Softmax
- Loss：Categorical_crossentropy
- Optimizer：adam
- Metrics：Accuracy

2. 將資料匯入模型中並進行訓練
3. 使用 Precision, Recall 及 F1 指標做類別預測，並用 Weighted-average 綜合多類別預測績效。
4. 輸出預測績效結果

表 8

Adult Dataset 多分類預測模型架構表

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 64)	3008
dense_2 (Dense)	(None, 32)	2080
dense_3 (Dense)	(None, 32)	1056
dropout_1 (Dropout)	(None, 32)	0
dense_4 (Dense)	(None, 16)	528
dense_5 (Dense)	(None, 16)	272
Total params: 6,944		
Total params: 6,944		
Non-trainable params: 0		

3.4 實驗結果

3.4.1 Shill Bidding Dataset

表 9

分類模型預測績效表

	訓練資料績效	驗證資料績效	測試資料績效
Accuracy	0.981	0.991	0.981
Loss	0.045	0.029	0.046

表 10

分類模型評估表

分類模型評估指標	
Precision	0.117
Recall	1.0
F1-score	0.210

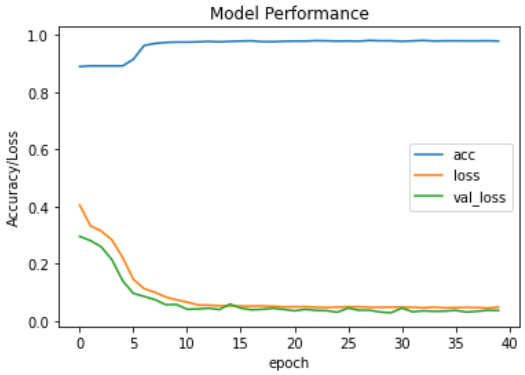

圖 4 分類模型訓練績效

表 11

迴歸模型預測績效表

	訓練資料績效	驗證資料績效
Loss	0.011	0.013

表 12

迴歸模型績效評估表

迴歸模型評估指標	
MAPE	7.29(%)
RMSE	0.108
MAE	0.052

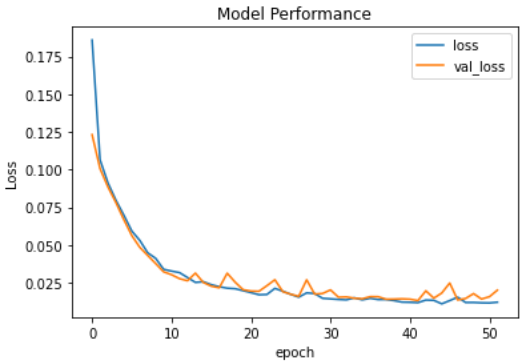

圖 5 迴歸模型訓練績效

3.4.2 Adult Dataset

表 13

分類模型預測績效表

	訓練資料績效	驗證資料績效	測試資料績效
Accuracy	0.433	0.427	0.426
Loss	1.646	1.691	1.683

表 14

分類模型評估表

分類模型評估指標	
Precision	0.355
Recall	0.426
F1-score	0.349

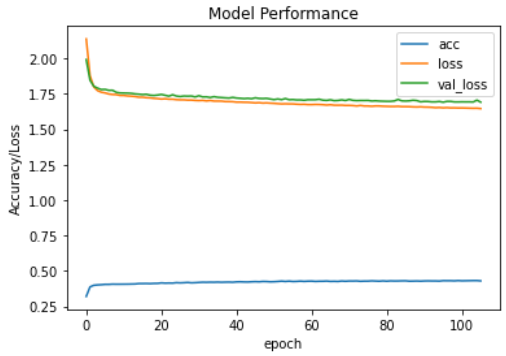

圖 6 分類模型訓練績效

四、 結論

在本研究使用 Shill Bidding DataSet 與 Adult Dataset，作為實驗資料集，在 Shill Bidding DataSet 進行兩項預測任務，分別為二分類預測以及數值預測。任務一主要希望透過神經網路之訓練模型來得知模型是否能在新的資料預測出是否是詐欺交易，以預防欺詐拍賣的發生，透過表 10 評估指標得知，Recall 較 Precision 高，代表能精準判定不正常交易的行為，而 Precision 較低，代表在預測出為不正常交易行為之中，實際為正常行為之比例，由實驗結果得知該分類模型能在未知資料中嚴謹抓到不正常的拍賣行為。任務二主要為預測拍賣者之得標率，在表 12 中 MAE、MAPE 與 RMSE 指標數值，在 MAE 與 RMSE 數值皆接近 0，在 MAPE 指標為 7.29%，代表該預測模型能預測出拍賣者之得標率。

在 Adult Dataset 中為收集之成年人口普查收入之相關數據，主要目的為預測成年人之教育程度，根據表 14，Precision、Recall 與 F1-score 指標皆小於 0.5，代表該模型未能準確地判斷該名成年人其教育程度為何。

參考文獻

英文文獻

Alzahrani, A., & Sadaoui, S. (2020). Clustering and labeling auction fraud data. *Data Management, Analytics and Innovation*, 269-283.