

國立雲林科技大學

資訊管理系所

資料探勘專案作業報告

學生/學號：

廖 昶 登 D10823001

張 靖 M10823001

劉博茜 M10823015

盧迺安 M10823023

指導教授：許中川

中華民國 108 年 11 月

摘要

本研究針對台灣客戶的拖欠付款情況，並比較了六種數據挖掘方法中拖欠概率的預測準確性。隨著大量資料的快速累積以及演算法與雲端運算的發展，大數據分析已經成為學術界與各行業關注與學習的焦點。從巨量資料中篩選有用的資訊，其運用的方法為資料探勘。本研究將使用 Python 撰寫最近鄰居分類器和決策數之程式進行模型訓練，並透過兩個分類器來比較，哪一個分類器在該資料集有比較好的績效。在 KNN 和決策樹訓練之下，可以得知 KNN 在信用卡欠款資料集的分類準確度，已收斂在 0.78。而決策樹的分類準確度則在 0.825。從績效上的角度來看，可以知道決策樹的分類在這筆資料集上，是比 KNN 來的要好。

關鍵字：資料探勘、Python、KNN、決策樹、信用卡

目錄

摘要	i
第 1 章 緒論	1
1.1 研究背景與動機	1
1.2 研究目的	1
第 2 章 方法	2
2.1 程式架構- KNN	2
2.2 程式架構-決策樹	2
第 3 章 實驗	4
4.1 資料集	4
4.2 實驗設計	4
4.3 實驗結果	4
第 4 章 結論	6

第1章 緒論

1.1 研究背景與動機

本研究針對台灣客戶的拖欠付款情況，並比較了六種數據挖掘方法中拖欠概率的預測準確性。信用卡則是於 1915 年起源於美國，最初的目的只是讓消費者憑卡購買物品、支付勞務費用和小額現金。但是近年來由於信用卡為人們的日常生活消費提供了許多便利，已發展成為一種極為普及的支付工具，可讓民眾享有「先享受，後付款」的服務。信用卡可在國際通用，其服務項目從購物、住宿、旅遊、醫療、保險及信用支付等，而銀行業也透過異業結盟提供各行各業多種優惠折扣，如：累積里程數、現金回饋、累積紅利點數以及分期付款零利率等。

1.2 研究目的

本研究主要以資料探勘為基礎，透過 Python 撰寫最近鄰居分類器和決策樹之程式，以此進行模型訓練，這項研究針對台灣客戶的拖欠付款情況，並比較了六種數據挖掘方法中拖欠概率的預測準確性。從風險管理的角度來看，估計違約概率的預測準確性的結果將比分類的二元結果(可信或不可信的客戶)更有價值。由於實際的違約概率未知，因此本研究提出了一種新穎的“排序平滑法”來估計違約的真實概率。將違約的真實概率作為響應變量(Y)，並將違約的預測概率作為自變量(X)，簡單線性回歸結果($Y = A + BX$)表明，人工神經網絡生成的預測模型 具有最高的確定係數；其回歸截距(A)接近零，回歸係數(B)接近1。因此，在六種數據挖掘技術中，人工神經網絡是唯一可以準確估計實際違約概率的技術。並透過兩個分類器來進行績效的比較。

第2章 方法

2.1 程式架構- KNN

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
import numpy as np
import matplotlib.pyplot as plt

data = pd.read_csv("D:\Python\DM\default_of_credit_card_clients.csv")
data = data.drop('ID', axis=1)
data_feature_name = data.columns[0:-1]
Y = data['default_payment_next_month']
X = data.drop('default_payment_next_month', axis=1)
X = data[data_feature_name]

X_train, X_test, y_train, y_test = train_test_split(X, Y, random_state=0,
train_size=0.8)
result = []
# 建立 KNN 分類器
for k in range(1,100):
    clf = KNeighborsClassifier(n_neighbors=k)
    credit_clf = clf.fit(X_train, y_train)
    acc = clf.score(X_test,y_test)
    result.append(acc)
    # print('K:%d Acc:%5f'%(k,acc))
result = np.array(result)
result = np.reshape(result,(-1,99))
plt.plot(np.mean(result,axis= 0))
```

2.2 程式架構-決策樹

```
import numpy as np
import pandas as pd
import pydotplus
from sklearn import tree
from sklearn.model_selection import train_test_split
```

```

import os
from sklearn.metrics import accuracy_score
data = pd.read_csv("D:\Python\DM\default_of_credit_card_clients.csv")
# data.head()
# print(data.head())
# data.isnull()
# data.columns

#提取訓練集與測試集
data = data.drop('ID', axis=1)
data_feature_name = data.columns[0:-1]
Y = data['default_payment_next_month']
X = data.drop('default_payment_next_month', axis=1)
X = data[data_feature_name]
X_train, X_test, y_train, y_test = train_test_split(X, Y, random_state=0,
train_size=0.8)

#建構決策數模型
model_tree = tree.DecisionTreeClassifier(max_depth=5)
model_tree.fit(X_train, y_train)

# 評論模型準確度
y_prob = model_tree.predict(X_test)
a = accuracy_score(y_test,y_prob)
print(a)

#可視化
data_ = pd.read_csv("D:\Python\DM\default_of_credit_card_clients.csv")
data_ = data_.drop('ID', axis=1)
data_feature_name = data_.columns[0:-1]
data_target_name = np.unique(data_["default_payment_next_month"])
os.environ["PATH"] += os.pathsep + 'C:/Program Files (x86)/Graphviz2.38/bin/'
graph =
tree.export_graphviz(model_tree,out_file=None,feature_names=data_feature_name,
                      class_names=['1','0'],filled=True,
rounded=True,special_characters=True,max_depth=5)
graph = pydotplus.graph_from_dot_data(graph)
graph.write_pdf('tree5.pdf')

```

第3章 實驗

4.1 資料集

表格 3.1 資料集

數據集特徵：	多變量	實例數：	30000	區域：	商業
屬性特徵：	整數，實數	屬性數量：	24	貢獻日期	2016-01-26
相關任務：	分類	缺少價值？	N/A	網頁點擊數：	405564

Attribute Information: 這項研究採用了一個二元變量，即默認付款(是= 1，否= 0)作為響應變量。這項研究回顧了文獻，並使用以下 23 個變量作為解釋變量：X1：給定信用額度（新台幣）：既包括個人消費者信用額度，也包括其家庭（補充）信用額度 X2：性別（1 = 男性； 2 = 女性）。X3：教育（1 = 研究生院； 2 = 大學； 3 = 高中； 4 = 其他）。X4：婚姻狀況（1 = 已婚； 2 = 單身； 3 = 其他）。X5：年齡（年）。X6-X11：過去的付款歷史。我們追蹤了過去的每月付款記錄（從 2005 年 4 月至 9 月），如下所示：X6 = 2005 年 9 月的還款狀態；X7 = 2005 年 8 月的還款狀態；X8 = 2005 年 7 月的還款狀態；X9 = 2005 年 6 月的還款狀態；X10 = 2005 年 5 月的還款狀態；X11 = 2005 年 4 月的還款狀態。還款狀態的度量標準為：-1 = 適當付款；1 = 付款延遲一個月；2 = 付款延遲兩個月；...；8 = 付款延遲八個月；9 = 付款延遲 9 個月以上。X12-X17：賬單金額（新台幣）。X12 = 2005 年 9 月的帳單金額；X13 = 2005 年 8 月的帳單金額；X14 = 2005 年 7 月的帳單金額；X15 = 2005 年 6 月的帳單金額；X16 = 2005 年 5 月的帳單金額；X17 = 2005 年 4 月的帳單金額。X18-X23：以前的付款金額（新台幣）。X18 = 2005 年 9 月支付的金額；X19 = 2005 年 8 月支付的金額；X20 = 2005 年 7 月支付的金額；X21 = 2005 年 6 月支付的金額；X22 = 2005 年 5 月支付的金額；X23 = 2005 年 4 月支付的金額。

4.2 實驗設計

利用 python 進行實驗，透過原始數據(training data)進行決策樹、和 KNN 的訓練，藉此建立該數據的分類模型，透過測試數據(testing data)進行結果的預測分類，透過準確值的比較來了解，兩者分類器績效的好壞。

4.3 實驗結果

利用 python 進行 KNN 和決策樹的程式撰寫，產生圖表，以下圖 4.1 和 4.2 分別是 KNN 的績效圖，和決策樹的展開圖。

第4章 結論

本研究的目的是利用 python 建構 KNN 和決策樹在信用卡欠款的分類模型，在 KNN 和決策樹訓練之下，可以得知 KNN 在信用卡欠款資料集的分類準確度，已收斂在 0.78。而決策樹的分類準確度則在 0.825。從績效上的角度來看，可以知道決策樹的分類在這筆資料集上，是比 KNN 來的要好。