經濟史導讀

R06323031張君瑋

1.What ia the main question raised in the paper?
2.Why should we care about it?
Answer1&2:

在實驗時,有時候希望觀察到的是不同類型的受試者對treatment的反應,但通常沒有自然且直覺的受試者類別供實驗者來區分,這時通常會使用mixture model來對受試者分類。

$$f(y_i) = \sum_{\tau=1}^{T} \pi_\tau f_\tau(y_i|\boldsymbol{x_i}'\boldsymbol{\beta})$$

mixture model是假設受試者先驗上有T個類別,來參加實驗的受試者為這T個類別中抽出,也就是說,來參加實驗受試者的參數x是由一個有T個類別的generative model所產生。式中的π即為各個類別發生的先驗機率,f為各類別的機率分布,π與f可以用expectation maximization algorithm估計出來,如此一來就可以知道受試者的由各類別產生的後驗機率,從而以maxima likelihood來把受試者歸類。

但此方法要估計的參數有T個分布的先驗機率與T個機率分布的參數,對樣本的數量要求較大,而實驗的驗本數通常不足以支持mixture model的參數估計。並且因為實驗通常重複數次,同個受試者若重複實驗數次,樣本會是以panel的形式存在,而mixture model對於panel data的處理非常複雜,因為不能把一個受試者不同次的樣本視為獨立,因此不只需要受試者產生的樣本,還需考慮受試者不同次實驗error term的機率分布,故作者提出一個對樣本數量要求較少且適合處理panel data的分類模型。

3.What is the author's answer?
4.How did the author get there?
Answer3&4:
作者提出的分類模如下:

先將每一個受試者的資料分開,分別對每一個受試者的資料做迴歸,得到係數。也就是說若現在有i個受試者,t次實驗的資料,就會有i次迴歸,每次迴歸的資料有t個樣本。如此一來會得到i組係數,再來可以使用任何機器學習的分類方法將這i組係數分成需要的組數。
作者使用的分類方法為CART,這是一種tree base的監督式學習模型,因為決策樹考慮的是不同的分類方法對資料混亂程度的減少(例如entropy),因此對樣本數的要求較少;但決策樹非常容易overfitting,因此可能需要考慮使用bootstrapping(bagging)來避免,也就是使用random forest。

這個方法不僅在樣本數上的要求較少,並且較mixture model容易處理panel data。