Adaptive Data Augmentation for Supervised Learning over Missing Data

郭庭沂
生醫電資所 碩一
no AI Tool

1 PROBLEM

在相關技術的蓬勃發展下，機器學習已經被廣泛用在許多不同的領域當中。在技術上而言，訓練模型的方法、參數的選擇固然重要，但訓練資料的品質重要性不亞於前者，所謂「Garbage-in, Garbage-out」，若訓練資料的品質不好的話，訓練出來的模型預測的結果也會不理想。而在許多可能造成資料品質較為不好的原因之中，缺失值的處理是資料科學家在訓練模型之前經常會遇到的問題，至今也有許多不同填補缺失值方法的研究，像是以平均值、迴歸方法等進行填補，然而其限制在於，由於訓練資料和目標資料的缺失模式不同，當訓練資料和目標資料分別進行填補可能會造成資料分布發散，也就是在機器學習領域中常見的資料集偏移的問題，進而降低了模型的正確預測能力。

因此為了解決上述問題，此研究提出了 DAGAN，同時使用兩個 GAN 進行自適應式的資料擴增，主要學習目標資料的缺失模式，並利用學習到的缺失模式在訓練資料上進行資料擴增，最後再用擴增的資料集重新訓練模型，使得模型可以更適用於目標資料，而不會因為上述之資料集偏移的問題導致模型表現不佳。

2 PRIOR WORK

現存的資料填補方法可以分為三大類，分別是基於統計方法、判別模型和生成模型。統計方法一般為最簡單的填補方法，在含有缺失值的位置用零、平均值或是中位數等方式進行填補，但僅透過簡單的統計分析進行填補可能和真實值的差異很大，資料品質上的提升有限。基於判別模型的則是透過觀察存在的數據進行迴歸或是隨機森林預測缺失值並進行插補。而生成模型則是基於 GAN 模型進行缺失值的處理。然而這些已經被提出的方法在完全隨機缺失的資料中有好的表現，但在比較複雜的缺失模式像是隨機缺失或非隨機缺失的資料中，表現就不是那麼理想。而在監督式學習中，訓練模型資料和目標資料的不同缺失模式的插補問題是以往沒有被考慮到的，而這也是此研究的重點之一。

在資料清洗上，也有一些和資料錯誤偵測並進行資料修補的相關研究，像是已經被發表的 GDR 和 SCARED 是依照整體資料的關係，針對類別值進行偵錯；Baran 是從多個偵錯模型中提取錯誤並進行未被標準化或是異常值的修復。然而，這些方法專注於資料修復而不是針對缺失值的處理進行設計的，但若結合此研究，將會對於缺失值方面的處理有更顯著的提升。

在資料擴增方面亦有相當多的方法，例如影像上進行裁切、旋轉、翻轉等等方式，也有團隊同樣使用 GAN 進行資料擴增，但針對關聯性資料的缺失值造成之雜訊偏移進行自適應式的資料擴增是比較少被注意到的，也是本研究的重點。

3 SOLUTION

在進行兩個 GAN 的訓練之前，先預定義一個遮罩向量 $m \in \{0,1\}^N$ 來表示資料中存在及不存在的值，1 設定為存在，0 設定為不存在，因此我們可以把每個 tuple 以下面的式子進行表示：

$$x = \psi(x_g, m) = x_g \odot m + NA \cdot (1 - m)$$

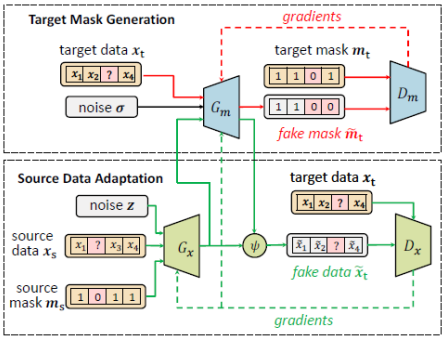
x_g 為沒有缺失值的 tuple，也就是真實值。此研究考慮了三種不同的缺失模式，分別是(1) 完全隨機缺失：m 和資料本身沒有關係；(2) 隨機缺失：m 和存在的資料有關係；(3) 非隨機缺失：m 和存在的資料以及缺失的資料都有關係。

而在 DAGAN 架構中的兩個 GAN 分別用於產生目標資料的缺失模式以及讓訓練資料依據前者產生出來的遮罩生成新的基於訓練資料的新資料，如圖(一)所示。在第一個負責產生目標資料缺失模式的 GAN 中，G_m 根據給的目標資料作為參照，將輸入的雜訊轉換生成一個假的遮罩並做為 D_m 的輸入，D_m 的工作是嘗試在 G_m 所生成出來的遮罩和真正的目標資料的遮罩中分辨異同，並回饋給 G_m，使得 G_m 能生成出越來越接近目標資料遮罩的假遮罩。

在第二個負責生成假資料的 GAN 中，同理 G_x 根據給的訓練資料和訓練資料的遮罩做為參照，將隨機產生的雜訊轉換生成一個假的資料，並結合上一個 GAN 所生成的假遮罩做為 D_x 的輸入，D_x 的工作則是嘗試在假資料和真資料中分辨出來，並同樣將結果回饋給 G_x，使得 G_x 可以生成越來越接近訓練資料的假資料。

經過兩個 GAN 之後我們可以得到新生成的假資料，而這些假資料保持了原來訓練資料不同屬性之間的關聯性，並且經過擴增的新資料集也和目標資料有一定的相似度。在這樣的結果之下，可以使用新的資料集進行簡單填補後重新進行

預測模型的訓練，使得新的預測模型能更好的預測目標資料的結果。



圖(一)： DAGAN 的架構

然而，問題是目標資料通常在訓練模型之前不會得到，所以無法提前得到目標資料的缺失模式。因此，此研究預先產生多種不同的缺失模式的資料，使得在訓練模型階段就可以依照不同目標資料的缺失模式進行訓練。每一種不同的缺失模式都可以重新訓練預測模型，得到許多符合不同缺失模式的預測模型，後續可根據目標資料的缺失模式選擇適合的預測模型使用。不過，若能訓練同一個預測模型使得在多種不同缺失模式的資料都可以使用的話將會方便許多，因此本研究也有另外將多種不同缺失模式所擴增的資料合併，做為一個綜合各種缺失模式得到的擴增資料的大資料集進行訓練，最後得到一個能適應各種缺失模式的預測模型。

4 RESULT

在進行各種不同缺失值填補方式的比較之前，首先要先決定要用哪些資料集做為實驗對象。此研究總共選擇了五種不同的資料集，其中三種為真實含有缺失值的資料集，另外選擇兩組未含有缺失值的資料。在實際實驗之前，分別對這兩組資料集進行完全隨機缺失、隨機缺失和非隨機缺失的處理，而每一種又根據訓練資料和目標資料缺失值之屬性有無重疊分為兩類，其中缺失率設定為 0.2 至 0.8 每間隔 0.1 產生一組，因此一種資料集會產生$3\times2\times7$共 42 種含有不同缺失值的新資料集，以便後續實驗分析不同缺失型態及缺失率在不同的填補方式下所訓練出來的分類器的表現。

為了評估此研究中所提出的 DAGAN 的表現，因此實驗同時也用 GAIN (使用 GAN 架構)、MISF (使用隨機森林)以及 MICE (使用分類及迴歸) 三種現有的填補方式做為比較的對象。

首先，以包含 DAGAN 在內的四種方式對預先準備好的三組真實含有缺失值的資料進行填補，並訓練分類器以測試集之結果計算 F1-score 做為其表現好壞的評估。在真實資料的實驗結果中發現 DAGAN 較其他種填補方法有更好的表現，可見避免訓練資料和目標資料之間產生的資料偏移使得在預測目標資料上有更好的表現。

另外針對兩組未含有缺失值的乾淨資料所產生的缺失值之屬性不重疊的資料集各 21 個同樣以四種方式進行填補並評估其表現。結果顯示在隨機缺失的資料集中，DAGAN 的 F1-score 平均來說分別高於 MICE、MISF 以及 GAIN 各 11.11%、31.79%以及 22.66%，在非隨機缺失中亦有相同的好表現。然而，在完全隨機缺失資料集中，DAGAN 並沒有明顯優於其他種填補方式的現象，推測可能是因為完全隨機缺失資料相對於其他兩者是較單純的缺失型態且其所造成資料偏移的狀況也相對不嚴重，再加上其他三種填補方式多是針對完全隨機缺失進行優化，因此 DAGAN 在表現上並沒有顯著的提升。不過在實際情況的應用上，大多缺失值還是以隨機缺失及非隨機缺失為大宗，因此 DAGAN 相對其他三種資料填補方式在實務上有其優勢。而缺失值之屬性重疊的資料集和上述討論的不重疊的結果相似，因此此處不另外提及。

而為了評估自適應性資料擴增(ADA)的效益，將進行 ADA 之後得到的資料集(AdaSrc)以及不含有缺失值的資料集(ClnSrc)分別進行分類器的訓練，並針對不同缺失率、不同缺失型態的目標資料進行預測。一個有趣的發現是，不論是在哪種缺失型態，ClnSrc 在缺失率比較小的資料集有更好的表現，AdaSrc 則是在缺失率比較高的資料集中有更好的表現。然而在缺失率小的資料集中，ClnSrc 的 F1-score 僅些微的高於 AdaSrc；在缺失率高的資料集中，AdaSrc 則是有 3%-18%的顯著提升，因此可推論 ADA 有其必要性。

透過實驗結果可以發現，和以往多種不同的資料填補方式相比，DAGAN 在不同的缺失型態及缺失率的填補上，更能有效的使得訓練之分類器在預測目標資料上有更好的結果。

5 CRITIQUE

雖然過往已經有研究使用 GAN 作為填補缺失值的方法，但其仍是將訓練資料及目標資料分開填補，意即沒有將資料偏移的可能性納入考量。而此研究特別針對資料偏移的現象提供了一個嶄新的方法，我認為利用 GAN 在訓練資料中加入目標資料的資訊達到資料擴增的效果，同時也提升分類器在目標資料上的預測能力，這樣的設計是十分有新穎性的。

6 POSSIBLE EXTENSION

由於此研究是建立在監督式學習上進行研究的，但眾所周知的是，將所有資料預先進行標註除了可能需要專家們的協助之外，若要有足夠的資料量進行模型的訓練，標註也會花上大量的時間。因此，若能將此方法同樣建立於半監督式學習甚至非監督式學習上，對於人力以及時間上的花費必將減輕不少，同時亦能有此研究所提供之更好的預測能力的結果。

除此之外，此研究先前提到的將所有不同缺失模式的資料合併訓練一個分類器和分別訓練不同分類器的實驗結果顯示，雖然用分布式穩健最佳化的技術進行合併訓練器的重新訓練，有時能較不同訓練器的效果為佳，但還是不同分類器表現較好的資料集比較多。因此或許可以嘗試其他最佳化的

技術，若能達到和不同分類器一樣的好表現的話，那麼在使用上相對會方便許多。

CITATION

[1] Tongyu Liu, Ju Fan, Yinqing Luo, Nan Tang, Guoliang Li, and Xiaoyong Du. 2021. Adaptive data augmentation for supervised learning over missing data. Proc. VLDB Endow. 14, 7 (March 2021), 1202–1214. https://doi.org/10.14778/3450980.3450989