

利用異常檢測與生成對抗網路進行數據合成

林鎔儒

國立中山大學應用數學所

摘要

在資料分析的應用中，已經有許多行之有年的數據合成方法，如合成少數類過採樣方法(Synthetic Minority Oversampling Technique，SMOTE)與自適應合成(Adaptive Synthetic，ADASYN)等。而近年來蓬勃發展生成對抗網路(Generative Adversarial Network，GAN)除了可用於生成圖片、影片、以及三維物體模型外，亦可生成一般結構化資料。而結構化資料生成結果與其他物件相比，生成品質較難以肉眼直接評斷，因此必須使用其他的檢驗方法來衡量生成樣本。

本研究主要採用 Wasserstein-GAN 建立一個數據合成的模型，並在其中結合一個異常檢測方法 (Anomaly Detection)，確保生產後的最低樣本品質，命名為 ADWGAN(Anomaly Detection in Wasserstein Generative Adversarial Network)。使用加州大學爾灣分校機器學習資料平台上的乳腺癌診斷資料集進行方法驗證，分別使用 SMOTE、ADASYN 與 ADWGAN 合成 357 個陽性樣本，並建立一高準確率的 Adaboost 分類器進行預測，其正確率分別為：0.986、0.945、1，顯示出 ADWGAN 的合成效果較前兩者佳。

關鍵詞：過採樣方法、SMOTE、ADASYN、GAN、WGAN、Anomaly Detection