合成資料（synthetic data）

「合成資料」（synthetic data）的出現，是為了保護原始資料所可能帶有的隱私資料或機敏資料，或是因法規或現實之限制而無法取得或利用研究所需資料的情況下，透過統計學方法、深度學習、或自然語言處理等方式，讓電腦以「模擬」方式生成研究所需之「合成資料」並進行後續研究跟利用，透過這個方法，資料科學家可以在無侵犯隱私的疑慮下，使合成資料所訓練出來的分類模型（classifiers）不會比原始資料所訓練出來的分類模型差。

在合成資料的生成技術當中，最熱門的研究為運用「生成對抗網路」（Generative Adversarial Network, GAN）形成合成資料（亦有其他生成合成資料之方法），生成對抗網路透過兩組類神經網路「生成網路」（generator）與辨識網路（discriminator）對於不同真偽目標值之反覆交錯訓練之結果，使其中一組類神經網路可生成與原始資料極度近似但又不完全一樣之資料，也就是具高度複雜性與擬真性而可供研究運用之「合成資料」。

英國國防科技實驗室（Defense Science and Technology Laboratory, DSTL）於2020年8月12日發布「合成資料」技術報告，此技術報告為DSTL委託英國航太系統公司（BAE Systems）的應用智慧實驗室（Applied Intelligence Labs, AI Labs）執行「後勤科技調查」（Logistics Technology Investigations, LTI）計畫下「資料科學與分析」主題的工作項目之一，探討在隱私考量下（privacy-preserving）「合成資料」當今技術發展情形，並提供評估技術之標準與方法。

技術報告中指出，資料的種類多元且面向廣泛，包含數字、分類資訊、文字與地理空間資訊等，針對不同資料種類所適用之生成技術均有所不同，也因此對於以監督式學習、非監督式學習或是統計學方法生成之「合成資料」需要採取不同的質化或量化方式進行技術評估；報告指出，目前尚未有一種可通用不同種類資料的合成資料生成技術或技術評估方法，建議應配合研究資料種類選取合適的生成技術與評估方法。

本文為「經濟部產業技術司科技專案成果」

相關連結

- Guidance: Synthetic Data (2020)
- 數位模擬分身（Digital Twin）
- 歐盟《歐洲資料戰略》

相關附件

- Pros and Cons of GAN Evaluation Measures (2018) [pdf]

→ 【南部場】113年「新創採購機制及鼓勵地方政府參與推動」說明會

→ 商業服務業個資行政檢查宣導說明會

范晏儒

專案經理 編譯整理

上稿時間：2020年10月

進階閱讀：

數位模擬分身（Digital Twin），資訊工業策進會科技法律研究所，https://stli.iii.org.tw/article-detail.aspx?tp=5&i=180&d=8257&no=67（最後瀏覽日：2020/09/07）。

歐盟《歐洲資料戰略》，資訊工業策進會科技法律研究所，https://stli.iii.org.tw/article-detail.aspx?tp=5&i=180&d=8433&no=67（最後瀏覽日：2020/09/07）。

文章標籤

個資去識別化 人工智慧 個人資料 資料開放 資料運用

推薦文章