經濟部 111 年度

《數位科技應用於產業發展計畫（2/4）》

合作研究計畫

《生醫資料商用智慧化工具-合成數據(Synthetic data)技術開發》

建議書徵求文件

財團法人資訊工業策進會

中華民國 111 年 1 月 3 日

111年度合作研究計畫建議書徵求文件

一、 簡介

生醫數據最常見的困難點是如何在不違反隱私保護規範的情形下，能將資料共享並進行分析應用。對於許多資料管理者而言，相關加密技術的應用門檻較高，因此若可應用一些技術來做到去識別化將有機會降低數據分享的門檻。去識別化技術包括加密(Encryption)、匿名化(Anonymization)、Access mediation/control、加噪(Adding noise)，包括置換加密Permutation、差分隱私(Differential Privacy)、機器學習(Machine Learning)，包括自動編碼器(Autoencoder)及聯邦學習(Federated Learning)。由於真實世界數據中，要取得帶有註釋的數據門檻較高，若可運用演算法創建合成數據，可用於增加數據的多樣性，提高研究的可重複性並保護隱私。

「數位科技應用於產業發展計畫」希望透過研發數據去識別化方法，產出不具備原始資料細節、但具備原始資料統計特性的資料集，來兼顧隱私保護與資料分析的需求，提高共享數據者的信心與意願，促進大數據資料產業能多元化的商用發展。

二、 計畫目標

本計畫以生醫資料的隱私保護與資料分析需求為目標，利用「合成數據技術」為基礎，運用目前相關結合差分隱私、深度學習等資料隱私去識別化技術，進行研究如何透過差分隱私技術、深度學習等相關技術，產出具備隱私保護與資料分析的需求的合成數據(Synthetic Data)模型，避免資料窺探者由去識別化數據來推估出原始資料。預期可達成目標如下：

- 原始數據(Raw Data)可透過差分隱私(Differential Privacy)技術與深度學習(Deep Learning) 技術等相關方法開發資料隱私去識別化演算法，用於產生合成數據(Synthetic Data)獲得相似結果。
- 原始數據(Raw Data)與合成數據(Synthetic Data)的 Inside test 與 Outside test的差異誤差小於10%。
- 原始數據(Raw Data)可透過分群(Clustering Classification)演算法進行目標分群，能同樣運用於合成數據(Synthetic Data)獲得相似分群結果。
- 生成數據技術必需能分別進行數值與離散型態的進行去識別化保護。
- 生成數據技術必需保有原始資料的分析相似性，能運作於數據欄位間關聯為線性關係與非線性關係的數據集。
- 可適當提供原始數據的前置處理(特徵工程)建議，促使合成數數據更符合生醫資料商用的應用。

三、 計畫範圍

所謂合成數據係指透過數學運算的方式，對原始資料集進行若干運算後，產生能保持大部分原始資料統計資訊的新數據集。而合成數據必須具有一定的隨機性，可以控制和調整合成數據背後的隨機過程，以控制資料隱私洩漏之風險。

而差分隱私(Differentail Privacy) 則是近十年來由Cynthia Dwork(時任微軟研究所)所提出的一種可以提供隱私保障的機制。基於差分隱私的原理，實際的系統開發者可以開發互動式或非互動式的隱私機制，而去識別化之合成數據的產出則屬於非互動式的機制。另外深度學習(Deep Learning) 則是透過將數據在關聯性的複雜度上進行降維以及合成數據產出，深度學習中有作用類似主成分分析(PCA，Principle Component Analysis)的架構，可以進行降維作業，也可以生成合成數據。而深度學習架構具有多項可調參數，可對架構進行調整，以了解合成數據在隱私與統計可用性上的變化。

綜觀現有針對合成數據的研究，希望能透過產出不具備原始資料細節、但具備原始資料統計特性的數據集，來兼顧隱私保護與資料分析的需求。總結，本研究計畫將現有相關研究結合差分隱私與深度學習等相關技術，用以開發資料隱私去識別化演算法，用於產生合成數據以保護資料的隱私性及保留資料的可用性。

四、 預期成果

項目	交付項目	交付內容	數量	交付型態	交付日期
1	期中報告	完成「合成數據(Synthetic data)技術開發」期中報告 1 份，內容涵蓋： ● 合成數據軟體規格 ● 演算法模型設計 ● 生成合成數據架構設計 ● 參考文獻	1 份	電子檔	111 年 5 月
2	期末報告 (包含演算法軟體雛型)	完成「合成數據(Synthetic data)技術開發」期末報告 1 份，內容涵蓋： ● 測試生成合成數據設計 ● 生成合成數據測試報告 ● 演算法軟體雛型 ● 為檢驗合成數據，提出可建立後驗機制的預設衡量相似度之常用模型	1 份	電子檔 (程式碼*)	111 年 8 月

***程式碼是以 Python 3.0 以上撰寫以及交付**。

※前述成果如有專利構想或專利申請產出時，需注意專利申請之新穎性(novelty)。因凡經公開發表之研發成果，如擬申請專利，須於公開發表後 6 個月內完成，前述成果如是以論文方式公開發表，將無法取得大陸與歐盟等國之專利。

五、 執行方式

(一) 為強化研究者與本項技術人員成果交流，於計畫執行期間，合作計畫執行單位每月辦理至少 1 次與資策會的計畫管控會議 (含技術人員階段性成果討論、教育訓練..等)

(二) 合作計畫執行單位需定期繳交期中報告(5 月底前)、期末報告(8 月底前)各 1 份。

六、計畫期程及預估計畫總經費

計畫執行區間： 自簽約用印完成日起 至 111 年 09 月 15 日

總經費：600,000元

七、 驗收標準

依本案需求說明書之交付項目、交付內容、交付形態及交付時限為驗收標準

1. 111 年 5 月 31 日前完成「合成數據(Synthetic data)技術開發」期中報告 1 份
2. 111 年 8 月 31 日前完成「合成數據(Synthetic data)技術開發」期末報告 1 份

八、 技術能力需求

研究者專業背景應含下列相關技術能力和專長，並以近年有相關研究經驗者為最佳。參與人員具合成數據(Synthetic data)機制規劃設計專業與知識應用能力。關於技術能力需具備差分隱私技術、深度學習技術、生成數據技術相關專長。