應用蟻群最佳化與模糊分群演算法於資料隱私保護之研究

邵敏華[1]　鄭仁豪[1]
[1]國立屏東科技大學資訊管理系
{ mhshao, m10056015 }@mail.npust.edu.tw

摘要

隨著個人隱私資料經常被使用於探勘技術中，如何合理的保護隱私資料，同時又能保證資料的可用性，成為目前資訊安全領域面臨的重大挑戰。匿名化是目前資料在公開發佈環境下實現隱私保護的主要技術之一，而 k-匿名(k-Anonymity)就是一種針對資料匿名化的隱私保護技術，能夠有效防止鏈結攻擊所造成的隱私揭露問題。從過去研究發現，許多方法處理上都是以減少資訊量的損失來建立 k-匿名，而使得資料品質遺失，因此為了確保處理後的資料仍具有分析研究的可用性，本研究提出了一個以蟻群最佳化為基礎於 k-匿名模型的方法，此方法以獲得良好資訊品質為目標，利用蟻群最佳化的穩健性和全域搜尋的優點，透過費洛蒙資訊的更新機制加速螞蟻在搜尋最佳解的執行過程，並結合模糊 C-means 分群法進行分群的改善，因此能有效減少偏離值的影響，達到降低資訊損失量，資訊品質與分群效率的提升，使得分群後的結果具有資料隱私保護能力，並兼顧了安全和資訊的效用。

關鍵詞：k-匿名、隱私保護、蟻群最佳化演算法、模糊 C-means 分群演算法、資料探勘、屬性。

1. 前言

隨著網際網路的蓬勃發展，資訊科技的各項技術與應用，使得企業的資料量不斷快速的累積增加，當企業面臨重要決策時，決策者要如何在這龐大的資料中取得有用資訊以幫助企業制定營運策略模式，這時資料的彙整與分析便相當重要[3]。但不可忽略的是，在這些資料當中可能隱含著企業交易客戶的基本資料，例如到購物中心選購日常生活用品或到醫院掛號看診，我們都會留下自身的基本資料，也因為網路的普及，資訊呈現高度透明化，這些隱密資料可能在相互傳遞的過程中存在著被有心人士竊取個資的可能風險，因此，對這些涉及個人隱私之基本資料採取預先防範的保護動作便成為相當重要的議題。

當資料在做特定用途時，必需經過隱私保護處理方能發佈，而 k-匿名(k-Anonymity)就是一個有效使用於保護個人資料不被他人識別(Identification)出來的技術[1][8][9]，主要是以分群演算法為基礎將資料依標準識別符號屬性(即可以識別個人身份的屬性欄位，例如年齡、性別、郵遞區號等)作為分群基準，將資料分為一群相似的群集後，再將每一相似群集分別進行泛化(Generalization)或抑制(Suppression)等處理方式將資料本身進行去識別化(De-identification)，為確保資料不被辨識則必需控制資料的揭露程度以滿足 k-匿名條件[1][7]，否則會破壞原先資料的完整性，因此，要做到資料的隱私保護與完整性，亦能提高在執行探勘上的效率，是當前研究所要追求的目標。

蟻群最佳化(Ant Colony Optimization，ACO)屬於萬用啟發式演算法的一種，隸屬於資料探勘技術，是藉由觀察自然界中螞蟻尋找食物的過程中，找尋出距離最短的路徑。從過去研究[1][7][8][9]中整理了近年來 k-匿名的各種方法，發現鮮少探討萬用啟發法則於 k-匿名上，為此本研究提出以 ACO 與 FCM 分群演算法結合於 k-匿名的方法，首先資料先經過 ACO 演算法執行群集運算，使用費洛蒙資訊作為資料在進行分群時機率計算之依據，依序作指派動作，然後在每一回合螞蟻找出現階段最佳解後，再將其作為 FCM 的初始群中心和初始分群數，藉由 FCM 的改善，以找尋出更佳的解，強化分群後的結果，使資訊品質能夠更好。一般當匿名 k 值越大，對資料隱私的保護效果會越好，被識別的機率等於 1/k，因此，個人資料就很難被辨識。

2. 文獻探討

2.1 k-匿名模型

k-匿名是由學者 Sweeney 於 2002 年所提出的一種隱私保護模型[8], [9]，其用途是保護資料隱私不被惡意攻擊者使用鏈結攻擊辨識以及有效解決標準識別符號(QI)揭露隱私問題的資料匿名化技術。

資料匿名化處理的原始資料，例如統計資料、醫療數據等，一般以資料表形式儲存，而且表中的每一筆記錄對應一個個人(包含多個屬性欄位)。這些屬性欄位可以分為三類，如下：

1. 唯一標識符號(Explicit Identifier)

可以唯一清楚標識單一個體的屬性，例如姓名、身份證字號、電話等；而為保護個人隱私資料安全，資料在發佈前會做加密或刪除的動作。

2. 標準識別符號(Quasi-Identifier，QI)

令表 T{P_1,…, P_n}為一個有限資料筆數(元素)的資料表，而屬性{P_1,…, P_n}為表 T 的有限屬性集合。假設這一資料表是一個有著 n 個屬性維度的資料表，且定義 N 為資料表中的資料總筆數。

給定一資料表 T{P_1,…, P_n}，而屬性的子集合{P_i,…, P_j}⊆{P_1,…, P_n}是表示{P_i,…, P_j}的所有元素屬性屬於{P_1,…, P_n}，且一筆資料的 t∈T，則 t{Pi,…, Pj}為{Pi,…, Pj}在 t 中的屬性值{Vi,…, Vj}；因此，T{P_i,…, P_j}可以表示表 T 在屬性{P_i,…, P_j}上的總集合。

3. 敏感屬性(Sensitive Attribute，SA)

指包含隱私資料的屬性，如圖 1(a)為一個原始客戶交易資料表，且每一筆記錄對應一個唯一的客戶姓名，其中{Name}為 Identifier，而{Race, Zip Code, Sex, Age}為 QI，{Goods}則為 SA。

(a) 未經匿名化之客戶交易資料表 T

	Identifier	Quasi-Identifier (QI)				
	Name	Race	Zip Code	Sex	Age	Goods
1	Mike	White	94130	Male	21	NB
2	Ivy	White	94138	Female	28	CD
3	Ben	Black	94140	Male	32	Book
4	Davy	Asian	94149	Male	39	Cup

(b) 匿名化後之客戶交易資料表 T* (k=2)

	Identifier	Quasi-Identifier (QI)				
	Name	Race	Zip Code	Sex	Age	Goods
1	Mike	White	9413*	Gender	[20, 30]	NB
2	Ivy	White	9413*	Gender	[20, 30]	CD
3	Ben	Person	9414*	Male	[30, 40]	Book
4	Davy	Person	9414*	Male	[30, 40]	Cup

圖 1 客戶交易資料表

接著，再以圖 1(a)舉例，假設在表 T 中的標準識別符號(QI)如果滿足 k-匿名條件時，表示在 QI 中的屬性集合的每一種組合必需在 T 表中至少出現 k 次。唯一標識符號(Identifier)指可以直接識別出個人真實身分的屬性(例如：姓名、電話、身分證字號)。標準識別符號(Quasi-Identifier，QI)指在公開資料表中，一種以上的欄位屬性在其他的公開資料表當中也有相同的屬性，而這些屬性是可以識別出個人的真實身份。圖 1(a)為未經過匿名化處理之客戶交易資料表，圖 1(b)為客戶交易資料表達 2-Anonymity 狀態(k=2)後的匿名之結果，匿名過後的資料表一般以 T*作表示之。

2.2 k-匿名之匿名技術

目前 k-匿名的研究主要集中在如何對原始資料表進行有效的匿名化，即實現匿名效果最好、資料可用性最高、且時間空間成本花費最小。資料匿名化一般採用兩種基本操作方式使得資料表滿足 k-匿名的條件，分別為泛化(Generalization)與抑制(Suppression)[8]，該技術不同於一般的扭曲、擾亂和隨機化等方法，它能夠保持發佈前後資料的真實性與一致性。

首先，泛化是對資料的屬性進行更廣義、抽象的描述，且不失原意。屬性的種類可以區分成類別型屬性與數值型屬性兩種，如圖 2(a)Race 為類別型屬性，可以(White→Person)進行泛化動作，而圖 2(b)Zip code 為數值型屬性，可以使用符號「*」進行泛化(94140→9414*→941**→*****)，因此，可以將 94140 與 9414*之間表示為 94140 $\leq_D$9414*，而為符合上述的泛化關係則必須滿足：(1) ∀ D_i, D_j, D_z ∈泛化域：$D_i \leq_D D_j$, $D_i \leq_D D_z \Rightarrow D_j \leq_D D_z \vee D_z \leq_D D_j$，(2)在所有的泛化域中，最高的泛化層級必定為單一節點。由此可知，泛化層級越高，資料匿名的程度越高，故可以明顯看出泛化是將底層的資料做模糊化，且泛化後的資料和原始的資料有相關性。

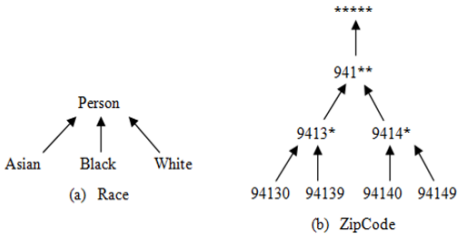

圖 2 屬性泛化層級示意圖

在 k-匿名化過程中，若某些記錄無法滿足 k-匿名條件時，一般採取抑制操作，係指在資料表或各類資料中，某些資料為了不被清楚辨識，採取的動作是將部份資料的相對應屬性利用其他符號作代替、隱藏或者刪除，亦即不發佈該屬性，以保持統計的特性，但由於抑制的轉換，隱藏的資料會有部份被破壞。

2.3 蟻群最佳化演算法

蟻群最佳化演算法(Ant Colony Optimization，ACO)是觀察真實世界的螞蟻覓食行為與特性將其轉換成數學模式，螞蟻在找尋食物的移動過程中會釋放出一種稱為"費洛蒙(Pheromone)"的化學物質，作為彼此溝通的工具[10][11][12][13]，因此其他的螞蟻便能間接得知食物位置，且可以此為依據來記錄行走的路徑，供後續抵達的螞蟻藉由殘留在路徑上的費洛蒙作為選擇移動路徑時的參考。

由於 ACO 是一個近似貪婪演算法的概念，亦

是一種找尋最佳解的方法，優點在不斷改善解字串(尋找周圍最佳解，然後往更佳的解作移動)，每次移動時都只選擇目前最佳的，直到無法改善為止。

1. 狀態轉換機率

在狀態轉換機率中，主要是利用開發(Exploitation)與探索(Exploration)兩種轉換機制來尋找路徑節點上要拜訪的下一個城市節點。

首先，開發指螞蟻依據目前最好的解來選擇下一個要走的城市，$u \in Jk(r)$表示 k 隻螞蟻在城市 i 所能選擇的城市 j (即已拜訪過的城市不能再重複拜訪)，如公式(1)；而 η 是一期望值，為 $1/d_{ij}$ 的倒數，q 則為一個介於 0~1 之間的隨機的參數設定值。

$$s=\begin{cases}\arg\max_{u\in Jk(r)}\{[\tau_{ij}(r,u)]\cdot[\eta_{ij}(r,u)]\}^{\beta} & ,\ if\ \ q\in q_0\\ S & ,\text{otherwise}\end{cases} \quad (1)$$

探索是指螞蟻會以隨機(Random)方式去挑選下一個要走的城市，可以增加解的廣度，如公式(2)；β 參數為控制城市之間關係程度之係數值。

$$S=P_k(r,s)=\begin{cases}\dfrac{[\tau_{ij}(r,s)]\cdot[\eta_{ij}(r,s)]^{\beta}}{\sum_{u\in Jk(r)}[\tau_{ij}(r,u)]\cdot[\eta_{ij}(r,u)]^{\beta}} & ,\ if\ \ s\in Jk(r)\\ 0 & ,otherwise\end{cases} \quad (2)$$

2. 費洛蒙更新

當螞蟻走過某一條路徑時會釋放費洛蒙在路徑上，而在路徑上的費洛的蒙濃度會隨著時間不斷更新揮發，其更新方式分為區域更新(Local Update)和全域更新(Global Update)兩種。區域更新指當每一隻螞蟻行經該路徑時會立即釋放費洛蒙作更新動作，即揮發與增加該路徑上的費洛蒙濃度。全域更新是指當全部螞蟻拜訪完所有城市後，每一隻螞蟻各自會產生一組解字串(Solution String)，將最佳的前幾組解依序作排序，更新最佳解路徑上的費洛蒙濃度矩陣，加速螞蟻在搜尋最佳解的執行過程。

2.4 模糊 C-means 分群法

模糊C-means分群法(Fuzzy C-means，FCM)是根據K-means分群演算法所衍生出來的模糊分群法，使用模糊邏輯(Fuzzy Logic)概念提升分群效果，由於具有"模糊"的特性，因此能利用隸屬函數(Membership function)使資料在不同情況下可以隸屬於不同的群集，不再絕對地屬於某一群集，而是以隸屬值(介於0~1)來表示資料隸屬於某一群集的程度[4]。在給定一個U上的一個模糊集A的定義為對任何$u\in U$都指定了一個數$u_A(u)\in[0, 1]$與之對應，因此$u_A(u)$稱為u對A的隸屬度，表示為$u_A: U \mid\rightarrow [0, 1]$且$u \mid\rightarrow u_A(u)$，這意味著此映射成為$A$的隸屬函數。從過去文獻知道，為分離出各資料所重疊的資訊，FCM是利用權重(Weight)方式給予模糊權重(Fuzzy weight)值，目的是希望能夠把偏離值的影響降到最低，以提高分群後的資料品質，因此FCM是有能力處理具重疊(Overlap)特性的資料[5]，這明顯會比傳統的資料分群法要來的合適且有效率。

計算目標函數值J_m，其計算如公式(3)。

$$J_m=\sum_{i=1}^{c}J_i=\sum_{i=1}^{c}\left(\sum_{j=1}^{n}u_{ij}^{m}\left\|X_j-C_i\right\|^2\right) \quad (3)$$

在執行FCM演算法時，需先設定初始參數，在各參數給定後，接著才開始計算初始群集中心，如公式(4)。

$$C_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}X_j}{\sum_{j=1}^{n}u_{ij}^{m}} \quad (4)$$

一般在計算目標函數值時，J_m值越小表示每一群集內的資料相似度越高，分群結果就越好；待J_m值計算好(數值最小)後，接著就可以計算每一筆資料與各群集之間的隸屬度，最後做FCM的模糊矩陣的更新動作，如公式(5)。

$$u_{ji}=\frac{1}{\sum_{k=1}^{c}\left(\dfrac{\left\|X_j-C_i\right\|}{\left\|X_j-C_k\right\|}\right)^{\frac{2}{m-1}}} \quad (5)$$

FCM的收斂條件是當$|J_m^{(t+1)}-J_m^{(t)}|<\varepsilon$符合終止條件時就停止，否則繼續執行計算，直至符合終止條件為止。

3. 研究方法

本研究根據研究問題，以 ACO 與 FCM 作結合來建構出本研究之演算法流程，也因為是用於 k-匿名，因此將其命名為 KFACO 分群演算法(k-Anonymity of Fuzzy C-means and Ant Colony Optimization)。我們依據螞蟻在資料點與群組的選擇、目標函數的計算，以及費洛蒙更新來執行資料的分群，利用 ACO 全域搜尋的分群處理特性，將大量資料劃分成數個群集，在 ACO 階段結束後，接著再藉由 FCM 進行分群結果的改善，提高在經過匿名化後的資料分群準確性，最後，透過費洛蒙資訊的更新機制，加速螞蟻在搜尋最佳解的執行過程，使匿名化後的資料擁有更好的資訊品質，降低資訊量損失。圖 3 為本研究之研究流程架構。

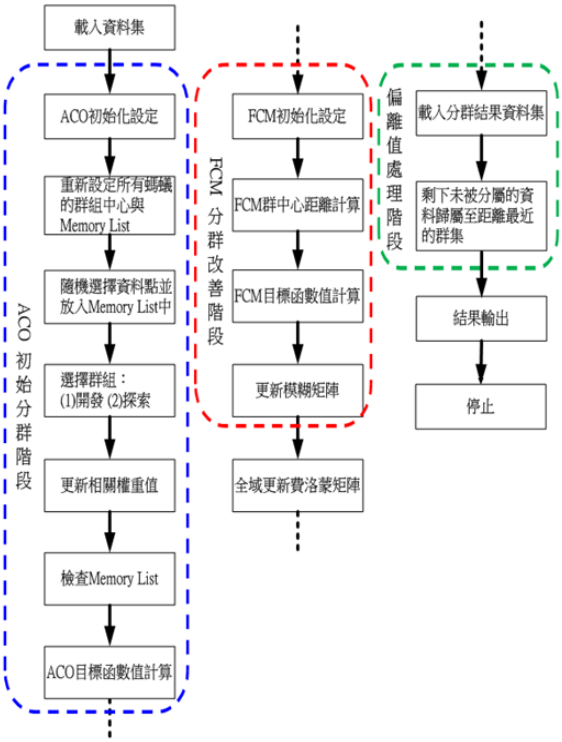

圖 3 研究流程架構

本演算法共分為三個階段，主要是以 ACO 為基礎，輔以 FCM 作結果的改善。

階段 1：ACO 初始分群階段

執行 ACO 全域搜尋作分群處理，結合 k-匿名隱私保護方法，使大量的資料不受偏離值影響，每一群集內的資料異質性最低，資訊量的損失降至最低，以求解出穩定的分群品質。

階段 2：FCM 分群改善階段

以 FCM 區域搜尋作階段 1 分群結果的強化，使用模糊權重於各群集中的資料，重新做計算，以改善資料的分群品質。

階段 3：偏離值處理階段

最後，於階段 3 處理剩下未被分群完整的偏離值(即雜訊)；最後做結果的輸出。

3.1 KFACO 分群演算法

本研究提出的KFACO分群演算法主要是利用 ACO 在全域搜尋上的優點，輔以 FCM 區域搜尋特性，來找出最佳分群數目及群集中心，使得求解出來的資訊品質能夠更好，也由於採用 FCM 中的模糊規則與權重進行推理，因此在穩健性與有效性上較好；KFACO 演算法的執行步驟如下：

步驟 1：測試資料集載入

將資料集依照 Identifier、QI、SA 三種屬性類別作劃分。將 Identifier 做刪除以保護原始資料，然後再篩選出 QI 做資料點轉換，轉換成 x 和 y 軸的數值型資料點，以便佈署於二維空間中。

步驟 2：ACO 初始參數設定

將費洛蒙濃度矩陣進行初始化，設定人工螞蟻數、菁英螞蟻數、疊代次數設為 0、門檻值 q_0 大小。

步驟 3：ACO 迴圈初始化

重新設定所有螞蟻的群中心與記憶表單(Memory List)。

步驟 4：資料點選擇

開始建構解(Solution)的次迴圈，在每個次迴圈執行時，每一隻螞蟻會以隨機的方式選擇下一個資料點，將已經選過的資料點置入螞蟻的記憶表單(已選過的資料點不能重複選取)中。

步驟 5：群組選擇

當一隻螞蟻所搭載的資料點 i 選擇群組時，有開發與探索兩種方式。選擇的方式是產生一隨機值 q(介於 0~1)，並與初始設定門檻值 q_0 作比較，如果 q 小於 q_0，則挑選現階段最佳的群組做配對；如果 q 大於 q_0，則依據 $P_k(r,s)$計算出機率值進行群組選擇。在群組選擇完畢之後，更新資料點 x_i 與第 j 群之間的關係權重值 W_{ij}。如果資料點 x_i 屬於第 j 群時，則 W_{ij} 設為 1；如果資料點 i 不屬於第 j 群的話，W_{ij} 則設為 0；相關公式如公式(6)及(7)所示。

$$W_{ij} = \begin{cases} 1, & \text{if data } i \text{ is clustered into cluster } j \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Minimize } J(W, C) = \sum_{i=1}^{m} J_i = \sum_{j=1}^{g} w_{ij} \left\| x_i - c_j \right\| \quad (6)$$

$$\text{where} \left\| X_i - C_j \right\| = \sqrt{\sum_{v=1}^{n} (x_{iv}, c_{jv})^2} \quad (7)$$

其中 g 為預設分群數、m 為資料點總數、n 為資料維度數目、X 是資料矩陣、C 則是群中心矩陣、W 代表權重矩陣、x_{iv} 為資料點 i 的第 v 個屬性值、c_{jv} 係指在第 j 群中的所有資料點的第 v 個平均屬性值。

步驟 6：檢查記憶表單

每一隻搭載資料點的螞蟻的記憶表單是否已填滿，若沒有則返回到步驟 3 重新執行，反之則繼續往下執行。

步驟 7：計算 ACO 目標函數值

所有螞蟻已經建構出完整解，且每一隻螞蟻會搭載一組解字串(Solution string)，而計算出來的目標函數值 J 可用來判斷哪一隻螞蟻的解字串是較佳的，依照大小作遞增(由小到大)排序，而第一個解為本次迴圈的最佳解，然後再將產生的最佳解當作 FCM 的初始輸入，用來做為分群結果的改善。

步驟 8：FCM 初始參數設定

承接 ACO 的次迴圈最佳解和目標函數值。設定權重係數 m 和終止條件 ε、模糊矩陣 U，疊代次數設為 1。

步驟 9：計算 FCM 的群中心和目標函數值

為了求得最佳的目標函數值 J_m，我們針對各輸入的參數進行微分計算，當 J_m 小於我們設定的容忍誤差時，停止疊代計算，否則返回步驟 7 重新執行。

步驟 10：更新模糊矩陣 U

待資料點與群集之間的群中心之隸屬程度計算出來後，立即更新模糊矩陣 U。此時如果

$|J_m^{(t+1)} - J_m^{(t)}| < \varepsilon$ 則停止計算，否則返回步驟 9 重新執行。

步驟 11：費洛蒙矩陣的更新

我們採取費洛蒙"全域更新"方式，如公式(8)。在 ACO 搭配 FCM 找到分群最佳解後，判斷如果找到的解比原先 ACO 的解要來的更好的話，就依照新的解作為費洛蒙更新之依據，否則就依照原來的解作費洛蒙更新動作。其中 t 表示時間(即迴圈次數)係數、ρ 為費洛蒙遺留下來的係數、$(1-\rho)$為費洛蒙揮發係數。

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \Delta\tau_{ij} \quad i = 1,\ldots,m \; ; \; j = 1,\ldots,g \tag{8}$$

$$\Delta\tau_{ij} = \begin{cases} \frac{M}{d_{ij}} & (v_i, v_j) \in path_{best}(t) \\ 0 & else \end{cases}$$

步驟 12：是否已達到最大迴圈次數

在更新完費洛蒙濃度矩陣後，檢查是否已達 ACO 最大迴圈次數，如未達到，則返回步驟 2 讓所有螞蟻重新執行出發，找尋下一個迴圈的解。如果已達最大迴圈次數，則停止該演算法，輸出目前的結果，此時所輸出的解就是本演算法所能找到最佳解的資料分群結果。

步驟 13：偏離值處理

在經過分群處理過後，我們將得到的分群結果中未被指派的偏離值依照距離平方公式計算歸屬至距離最近的群集之中，最後才是我們所要的分群結果，然後才做輸出動作。

3.2 資訊損失量之量測

學者 Byun 於 2007 年提出一種用來度量資料在匿名化過程中，資料所損失的量值的計算方法[6]，如公式(9)。

$$D(G_k) = \sum_{i=1}^{m} \frac{Max_i(G_k) - Min_i(G_k)}{Max_i(T) - Min_i(T)} + \sum_{j=1}^{q} \frac{H(T_{c_j}(G_k))}{H(T_{c_j})} \tag{9}$$

在本研究中我們加入了權重的概念以作為計算各個資料點對每一群集的資訊損失量為基礎，如公式(10)。

$$D(G_k) = \sum_{i=1}^{m} \left(\frac{Max_i(G_k) - Min_i(G_k)}{Max_i(T) - Min_i(T)} \right) * W_{m_i} + \sum_{j=1}^{q} \left(\frac{H(T_{c_j}(G_k))}{H(T_{c_j})} \right) * W_{q_j} \tag{10}$$

在 $WG(G_k)$的 $Max_i(G_k)$、$Min_i(G_k)$、$Max_i(T)$、$Min_i(T)$、$H(T)$ 等因子跟原本 $D(G_k)$ 的因子一樣不變時，Wm_i 跟 Wq_i 表示使用者對該欄位所給予的權重值，而整個資料表的總體資訊損失量的計算如公式(11)。

$$IL(T) = \sum_{k=1}^{g} IL(G_k) \tag{11}$$

	Zip Code	Age	Gender
G_1	12713	21	Male
G_1	12451	23	Female
G_1	12411	34	Female
G_2	12736	39	Male
G_2	12713	30	Female
G_2	12799	20	Male

圖 4 資訊損失量計算之範例資料表

舉例來說，圖 4 的欄位給予的權重值分別為 *ZipCode*=3、*Age*=2、*Gender*=1，如果圖 4 當中的前三筆資料為資料集 G_1，而後三筆資料為資料集 G_2，則資料集 G_1 的資訊損失量計算結果為：$IL(G_1)=3*(\frac{34-21}{39-20}*2+\frac{2}{2}*3+\frac{1}{1}*1)=16.105$，而資料集 G_2 的資訊損失量計算結果則為：$IL(G_2)=3*(\frac{39-20}{39-20}*2+\frac{1}{2}*3+\frac{1}{1}*1)=13.5$。因此，我們在經由上述計算，可以得到整個圖 4 資料集的總體資訊損失量為 $IL(T)$= 16.105 + 13.5 = 29.605，而資料集 G_2 的資訊損失量要比資料集 G_1 的資訊損失量來的較小。因此，本研究的目標函數值為求得每一個群集的權重資訊損失量最小化，也就是 *Min*(*Function*)，即 *Function*=$L(G_k)$。

3.3 偏離值的處理

在經過階段 1 與階段 2 分群計算過後，各個資料點都已經被歸屬至某個群集之中，假如所有群集的大小皆已經滿足 k 筆資料，則將剩餘資料，也就是偏離值(Outlier)依照距離平方公式計算，將其計算結果指派至最接近的群集之中，這樣可以不必受群集大小的限制，便可以得到一個完整的群組集合，且所包含在內的任一群集皆已經滿足 k-匿名的條件限制，並且具有良好的資料分群品質。

4. 實驗設計

4.1 開發工具與實驗平台

本研究在開發工具上以 Microsoft Visual Studio 2010 for C#作編寫，以 Intel Core(TM) i5-2500 CPU@3.30GHz 之中央處理器與 4GB 之記憶體為實驗平台，作業系統為 Windows 7。

4.2 測試資料

本研究採用國際通用資料庫，資料來源為 UCI repository of machine learning databases 網站上[2]所下載的 Adult dataset，是一般用來評估 k-匿名結果的公開資料集，在刪除遺漏值不採計部分，總共剩有 30,162 筆資料做為本實驗測試之用；相關資訊如圖 5 所示。

Data Set Characteristics：	*Multivariate*	Number of Instances：	*48842*
Attribute Characteristics：	*Categorical, Integer*	Number of Attributes：	*14*
Associated Tasks：	*Classification*	Area：	*Social*

圖 5 Adult dataset

4.3 資料預處理

將實驗的 Adult dataset 載入，依照使用者自訂方式對其做資料前置處理，我們將屬性欄位劃分為 QI、Identifier、SA 三種屬性類別，如圖 6 所示。

資料行名稱	資料類型	允許 Null
aid	int	☑
age	int	☑
education	int	☑
workclass	nvarchar(50)	☑
maritalstatu	nvarchar(50)	☑
occuption	nvarchar(50)	☑
race	nvarchar(50)	☑
gender	nvarchar(50)	☑
nativecountry	nvarchar(50)	☑
salary	nvarchar(50)	☑
x	nvarchar(50)	☑
y	nvarchar(50)	☑
		☐

圖 6 Adult dataset 欄位設定

在執行 k-匿名資料保護之前，我們必須先觀察資料集中的每一個屬性欄位，以了解存於資料集的屬性是數值型(Continuous)屬性資料還是類別型(Categorical)屬性資料，而圖 7 為我們將 Adult dataset 載入 SQL Server 後，資料經過維度轉換處理後，已經從真實資料轉換成 x 軸和 y 軸的數值型資料點型態，並且佈署於二維的空間之中，方便做分析。

	aid	age	education	workclass	maritalstatu	occuption	race	gender	nativecountry	salary	x	y
1	25848	25	10	Private	Separated	Adm-clerical						
2	25851	35	10	Private	Married-civ-spouse	Handlers-cleaners						
3	25865	32	10				White	Female	United-States	<=50K	-0.041941	-0.040405
4	25879	20	10				White	Male	United-States	<=50K	0.020246	0.20513
5	25880	27	10				White	Male	United-States	<=50K	0.094149	0.15186
6	25883	34	10				White	Female	United-States	<=50K	0.012343	-0.0026182
7	25885	24	10				Black	Male	United-States	<=50K	-0.054915	-0.034662
8	25887	20	10				Black	Male	United-States	>50K	-0.02132	0.17828
9	25888	51	10				White	Male	United-States	>50K	-0.0035234	-0.01026
10	25892	19	10				White	Female	United-States	<=50K	0.0084819	-0.0012812
							White	Male	United-States	<=50K	0.033219	0.032732
							Black	Male	United-States	<=50K	0.024896	0.0019971

圖 7 資料維度轉換

4.4 實驗分析

本研究主要是以 k-匿名隱私保護為主，在實驗驗證部份將會採用現有 k-匿名的隱私保護資料分群演算法作為本研究的實驗對照組。在比較的項目上主要以資訊損失量總和及分群品質為主，如下：

1. 總體資訊損失量

在本實驗階段因為對照組和本研究的方法在資訊損失量的計算公式不一定相同，在比較時會針對實驗對照組的資訊損失量計算方式，與本研究的方法之分群結果下去計算，方便進行比較；而本研究的資訊損失量計算方式也亦會用於對照組的分群結果下去做計算並進行比較。

2. 資訊品質

依照資料分群在各階段的處理方式的不同，將執行的 k-匿名分群結果進行比較，針對群集之間的變異程度做比較，以便做分群品質的評估。

5. 結論

本研究基於以資料的隱私保護為前提下，為了維持資料在做探勘時的可用性，故設計了以螞蟻覓食時的特性為基礎，讓資料集當中的資料欄位的屬性資料能夠依據 k-匿名理論進行資料分群的動作，同時還要符合隱私保護的基本要求。

在現階段我們將先以小樣本模擬演算法之執行過程，未來在實驗的驗證部分，將針對 KFACO 演算法所建構出的資料分群結果來計算資訊損失量、執行時間與不同參數下之執行結果。而因為對照組和本研究的方法在資訊損失量的計算公式不一定相同，所以在比較時會將對照組的資訊損失量計算方式針對本研究的方法之分群結果下去計算，進行比較；而本研究的資訊損失量計算方式也亦會用於對照組的分群結果下去做計算並進行比較。在資訊品質方面會依據系統執行 k-匿名所分群出來之分群結果好壞與對照組做相互的比較。

總體而言，本研究由於結合具有全域性和穩健性優點之 ACO 演算法與具有優良分群品質及能夠處理具重疊性資料的 FCM 分群法兩者，在分群結果上將能夠達到改善資料解品質的效果，使得分群效率進而提升，降低資訊損失量，在提升高資訊效用的同時亦能夠達到保護個體隱私權益之目的。

參考文獻

[1] A. Friedman, R. Wolff and A. Schuster, "Providing k-Anonymity in Data Mining," in the VLDB Journal, Israel, 2007, pp. 789-804.

[2] C. Blake and C. Merz. Uci Repository of Machine Learning Databases. 1998.

[3] C. Legany, S. Juhasz and A. Babos, "Cluster Validity Measurement Techniques," the 6th International Symposium of Hungarian Researchers on Computational Intelligence, 2006, pp. 388-393.

[4] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms," Plenum, 1981.

[5] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," Journal of Cybernetics, 1973, pp. 32-57.

[6] J. W. Byun, A. Kamra, E. Bertino and N. Li, "Efficient K-Anonymization Using Clustering

Techniques," in Internal Conference on Tabase Systems for Advanced Applications, USA, 2007.
[7] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," in Proceedings of the 22nd IEEE International Conference on Data Engineering, USA, 2006.
[8] L. Sweeney, "Achieving K-Anonymity Privacy Protection using Generalization and Suppression," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, USA, 2002.
[9] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, USA, 2002, pp. 571-588.
[10] M. Dorigo, V. Maniezzo and A. Colorni, "Ant System: An Autocatalytic Optimizing Process," Technical Report No.91-016, Italy, 1991.
[11] M. Dorigo and L. M. Gambardella, "Ant colony system: a cooperative learning approach to the traveling salesman problem," IEEE Transactions on Evolutionary Computation, 1997, pp. 53-66.
[12] O. A. Mohamed Jafar and R. Sivakumar, "Ant-based Clustering Algorithms: A Brief Survey," International Journal of Computer Theory and Engineering, 2010, pp. 1793-8201.
[13] P. S. Shelokar, V. K. Jayaraman and B. D. Kulkarni, "An ant colony approach for clustering," Analytica Chimica Acta, India, 2003, pp. 187-195.