一個針對微聚合問題的分裂重組法

林惠珍
國立暨南國際大學
s97213519@ncnu.edu.tw

尹邦嚴
國立暨南國際大學
pyyin@ncnu.edu.tw

***摘要*—資訊科技的發達使得收集大量資料不再是件難事，然而這項利器卻引發了資料中悠關個人隱私的爭議。對於公開的統計資料庫來說，如何有效保護個人資料，除了制定相關法律保障之外，可以利用統計洩露控制技術的微聚合方法來對個人資料加以保護。微聚合方法的程序與傳統的分群法本質相似，且在過去的文獻中，以傳統分群法改良而成的微聚合方法佔有不少比例，可見其重要性。本研究針對此研究方向再提出兩個不同的方法，一個是由 c-means 分群法改良而成，另一個則是結合 c-means 與階層式分群法修改而來。我們希望融合兩者，截長補短，改善單一分群法為基礎的不足。在實驗中也印證了我們的想法，結合兩種方法的確能獲得單一方法所無法達到的效能。**

***關鍵詞*— 微聚合、統計洩露控制、資訊損失、c-means 演算法、階層式聚合演算法**

一、 前言

隨著時代的進步，資訊科技愈來愈發達，我們能夠透過電腦、網路等等科技工具輕易地收集到大量的資訊。而政府的統計部門以及許多的民間企業，根據不同的目的去收集大眾資料並進行分析，例如政府為了制定某項政策，想了解國家人口分佈情形或是欲對目前的國家經濟發展作分析；在民間企業中，則是需要對於客戶進行消費行為的分析，以利商品銷售等等。在收集到民眾或客戶們的資料之後，開始針對資料進行資料探勘(Data mining)、製成統計數據以及分析，其結果將成為政府以及各個企業單位決策時的參考依據。而後，這些資料通常會再被公開出去，以便其他人也能依其研究目的所需自由取用，被發佈的資料大致上有兩種類型：「微資料(micro-data sets，或 individual respondent records，即是民眾的個人資料，或是各企業單位、公司的營業資料)」和「表格資料(tabular data，聚集微資料後的統計資料)」[7]。

但是，發佈這些統計資料以及大量微資料給外界人士參考，雖然出發點是善意，有利於進行規劃與決策，但另一方面卻因這些資料中可能含有資料提供者個人隱私的資訊，可能引發危害個人隱私權的疑慮，假設資料提供者的身分跟著曝光，有些機密的資訊，例如個人財務情形、身體健康狀況等等就會被他人得知，造成當事者相當大的困擾，若讓有心人士得知，甚至有可能會對資料提供者進行犯罪與加害。因此，為了確保大眾的隱私權，在公開資料之前，必須對資料採取適當的保護措施，而這樣的保護機制稱作「統計洩露控制(Statistical Disclosure Control，SDC；或是Statistical Disclosure Limitation，SDL)」。

除了設立保護機制之外，在相關的法律條文中也必須保障民眾的隱私權，例如美國有「機密資訊保護及統計效率法 (Confidential Information Protection and Statistical Efficiency Act，CIPSEA)」，而台灣則是在1995年時通過立法程序制定了「電腦處理個人資料保護法」。有了法律條文的保障，民眾在提供資料時，才能安心地提供正確的資料，以確保後續資料探勘結果的正確性[6]。

SDC 機制大致上被應用於幾種主要的資訊隱私保護類型中[6]，第一類是針對表格型態的資料作保護，在這類型的方法中，其中一種是對於表格中敏感性與非敏感性的資料值分別做些不同程度的更動，但保持最後統計值(例如：平均、加總等)不變，如此，使得敏感性的原始資料值能夠有所掩飾而達到保密目的，又能維持資料統

計值的可用性，其技術稱作「表格調整控制(Controlled Tabular Adjustment ，CTA)」[3][11]；第二類是當有人進行動態資料庫查詢時，必須針對查詢者下達查詢指令後，資料庫所給予的回應資料作保護，以防發生查詢者從一系列的回應資料當中，分析、推斷出某些秘密資訊；第三類是對微資料作保護，此種保護技術稱作「微聚合(Microaggregation)」，藉由先將保護的資料作分群，再以每一群的代表值(例如：平均值)替換群中各個原始資料值，也就是以擾亂原始值的方式達到保護原始資料的目的。

然而，雖然要對資料採取保護措施，但也不能給予過多的保護，我們必須考慮資料經過保護過後的「可利用性(data utility)」。通常，在資料保護的過程中都會對資料產生不同程度的干擾，藉由改變其原始值使外人無法直接得知以達到保密性，改變得愈多則保密性愈高。但這樣一來，改變後的資料與原始值差異愈大，造成高資訊損失量(information loss)，將使得資料失去可利用性，那麼若針對這樣的資料進行資料探勘，其結果非常有可能不足採信；反之，若給予的保護太少，雖然留住了資料的可利用性，卻也因保密性不夠，而有可能造成個人資料外洩。因此，我們面臨的問題是，如何在「保密」與「資訊損失量」之間取得平衡點，也就是說，我們必須要同時達到兩個目標，一個是「維護隱私」，另一個則是「維護資料的可用性」。

本研究針對微聚合這部分的技術做深入的探討，嘗試設計出一種新的微聚合技術方法，來對微資料作保護。後續之文章內容編排如下：第二章介紹微聚合的基本原理，包含問題定義以及前人所作相關研究；第三章則闡述本研究提出的解決微聚合問題之方法；第四章進行實驗的介紹，並且將數據結果與前人所作實驗結果相比較；第五章則對本研究綜觀探討其貢獻及提出一些結論與未來可再發展的研究方向。

二、 相關文獻探討

(一) 微聚合定義與相關理論

微聚合是 SDC 中的一種技術，符合 k 匿名精神(k-anonymity)[2][10]，背後的原理是：先將所有資料分群，然後以群心取代該組成員的資料，藉由確定每一群中至少含有 k 筆資料，可以達到保護個別資料個體之隱私、匿名性的目標。因為每一群中存在至少 k 筆一樣的資料值，若是要從這些一樣的資料中識別出某個特定的資料個體，將會非常困難，如此一來，可避免洩露資料提供者的身份，微聚合的參數 k 可視為安全強度。

微聚合也可以被視為一種分群問題，只是它有別於傳統的分群，必須有群規模(group size)的限制，每一群至少都要包含 k 個資料元素。然而，我們在解微聚合問題時，仍然可以將傳統分群法加以修改，加入群規模的限制，即成為解決微聚合問題的方法之一，例如在 Domingo-Ferrer 和 Mateo-Sanz 的研究[7]中，即是將階層式分群法修改成為解決微聚合問題的演算法。

一般而言，在將資料作分群時，目標為最小化群內變異數，也就是說每一群內的資料同質性要高。相關研究普遍使用組內均方差總和(sum of squared errors，SSE)公式來衡量群內的資料同質性[12][14][24][25]，若 SSE 值小，就代表群內同質性高，相反地，SSE 大，則同質性低。SSE 公式如數學式(1)所示：

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i) \qquad (1)$$

g 代表群數，n_i 表示第 i 群內的資料總數，x_{ij} 表示第 i 群中的第 j 個資料點，$\bar{x}_i$ 則代表第 i 群的群心值(資料點的平均值)。另外，我們在微聚合問題中欲尋找最佳的分群結果 k-分群，而如果以 SSE 的角度來看，「最佳的 k-分群」可以定義成

「具有最小 SSE 值的 k-分群」[7]。

最小化 SSE，除了求得群內高同質性之外，其實還可以達到另一個目的，那就是最小化資訊損失。由於資訊損失是因為聚合程序中以群心替換個別資料時所造成的，藉由最小化 SSE，可以達到群內同質性高，而同質性高就代表群心與個別資料的差異不會太大，如此在替換之後所產生的資訊損失也就可以很小，所以說最小化 SSE，等同於最小化資訊損失。而前述觀念也能透過公式來解釋，衡量資訊損失的公式如數學式(2)所示：

$$L = \frac{SSE}{SST} \tag{2}$$

L 是資訊損失指標值，而 SST 是當所有資料都屬於同一群時所計算出來的 SSE 值，公式如數學(3)所示：

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})'(x_i - \bar{x}) \tag{3}$$

n 代表整個資料集共有 n 個資料點，x_i 代表第 i 個資料點，$\bar{x}$ 則是整個資料集的平均資料點。實際上，L 就只是正規化後的 SSE，值介於 0 到 1 之間，所以說最小化 SSE 其實就等同於最小化資訊損失的指標 L，而此一數值也成為微聚合問題中普遍被用來比較、衡量演算法之優劣的指標值[6][7][17][22]。

2002 年，Domingo-Ferrer 和 Mateo-Sanz 的研究[7]證明出最佳的 k-分群應該具有的群規模範圍在 k 至 2k-1 個資料元素，也就是說，如果 k-分群裡的每一群所包含的資料數目皆能在 k 到 2k-1 個的範圍裡，那麼將可以找到具有最小 SSE 值的最佳 k-分群。接著，我們可以藉由這個證明，把微聚合問題定義如下：

目標式：最小化 SSE

限制式：

$$k \le n_i \le 2k-1, \text{ for i = 1, 2,…,g} \quad , \tag{4}$$

$$\sum_{i=1}^{g} n_i = n \tag{5}$$

針對微聚合問題的相關研究有許多，除了群規模變動性的不同之外，另一個特性是資料的變量數目，分為「單一變量(Univariate)」與「多變量(Multivariate)」，單一變量資料集指的是一維的資料集，也就是資料集裡頭的每一筆資料只含有一個屬性；而多變量資料集則是指每一筆資料含有多個屬性，也就是一個多維度的資料集。單一變量資料集的微聚合問題會比多變量資料集的問題容易，因為資料只有一個維度，所以在分群的階段，要找到相似的資料只要透過此一維度的資料排序，由資料的前後順序便可以得知緊臨的資料就是較為相似的，應該分為同一群，在 Hansen 和 Mukherjee 的研究中[15]，即提出一個在多項式時間內就能解出最佳單一變量微聚合問題的方法；然而，多變量資料集的微聚合問題相對地較為複雜，在 Oganian 和 Domingo-Ferrer 的研究中[21]，證明對於多維度的資料集來說，欲解出最佳的微聚合是一個 NP-hard 問題，因此當我們處理多維度資料集的微聚合問題時，可採用啟發式的演算法(Heuristics)[6][7][17][22]，本研究也是以多變量資料集進行研究。

(二) 微聚合方法

過去有不少文獻提出不同的微聚合方法以改善資訊損失量，我們觀察這些方法的特性，在此以三種類型進行分類整理。

- Clustering-based

如同我們在前文的敘述，微聚合問題與分群問題在本質上是雷同的，所以直接將傳統分群法改制成為微聚合方法是一個相當可行的方式，只需要在群組合併、分裂的過程中限制群組規模即可。傳統分群法分為三大類：(1)分割式分群法(Partitioning)，最具代表性的即為「c-means(又名

k-means[1])」[13][18]，雖然c-means演算法能夠幫助我們快速求得解，但其分群之結果好壞容易受到一開始挑選的初始群心所影響。(2)階層式分群(Hierarchy)[13][24]，主要是依據資料點之間的相似度或是資料點與群心的距離將資料集以樹狀結構呈現出來，再藉由一連串循序分割或聚合的動作將資料分群，因此這類方法又可再細分為由上而下的「分裂法(Divisive approach)」與由下而上的「聚合法(Agglomerative approach)」兩類。(3)圖形式分群(Graph-based)，主要是將所有的資料點形成一個圖形(Graph)，再藉由刪掉邊來形成群，以最小擴張樹(Minimum spanning tree)作分群法就是一個例子。

- Distance-based

(1)最大距離法(Maximum distance method，MD)，Domingo-Ferrer 和 Mateo-Sanz 等人的研究[7][8]提出 MD 法，在文獻[17]中稱為「Diameter-based fixed-size，DBFS」方法，其方法說明如下：

A. 首先找到資料集當中距離(例如歐幾里德距離)最遠的兩個資料點 x_r 與 x_s 作為參考點形成兩個群組。以 x_r 為例，找到一個離 x_r 最近的點加入 x_r 這個群組，計算其群心後，再找到一個離此群心最近的點加入，反覆此流程，直到加入 k-1 個資料點，形成規模為 k 的群組；x_s 的群組也是一樣的作法。

B. 檢視資料集當中尚未被分群的資料點數量，如果還有 2k 個以上資料點，則再回到步驟 A.繼續分群；若是剩下 k 至 2k-1 個資料點，就將這些資料點獨立成一個群組；若是少於 k 個資料點，則將這些資料點加入離其最近的群組中。

(2)群心最大距離法(Maximum distance to average vector method，MDAV)，最早於文獻[16]提出，在文獻[6][22]也有提及，其方法流程與 MD 類似，其差異在步驟 A.找尋兩個資料點的作法為：計算出資料集的群心，找到離群心距離最遠的一個資料點 x_r，再找到離 x_r 最遠的一個資料點 x_s。(3)群心為基固定規模法(Centroid-based fixed-size method，CBFS)，學者 Laszlo 和 Mukherjee 的研究[17]中，將 MDAV 的方法稍作更改，以「CBFS」的名稱再度提出[6][22]，其主要差異為：CBFS 只找離群心距離最遠的一個資料點 x_r 並形成一個群組，不再另找 x_s。於是，MDAV 一次可以形成兩個群組，而 CBFS 只會形成一個群組。

- Genetic Algorithm

基因演算法(Genetic algorithm，GA)是一種生物演算法，在過去有學者將其運用於解分群問題中[20]，也由於分群問題與微聚合問題的相似性，從 GA 衍生出微聚合方法也是個可行的方式。這一類的方法在資料點的分群上需要考慮到 GA 染色體編碼的方式，一條染色體代表資料集的一種分群結果，一個基因代表一個資料點，而每個基因內的值即是資料點所屬的群組編號。

除了上述三大類微聚合的方法以外，有些學者提出另一種補強的方法，他們採二階段式，第一階段可以使用任何前文提及的微聚合方法，第二階段則是就第一階段產生的分群結果再次改善，其目的在於降低資訊損失量。這類型的方法彈性大、可用性高，因為第二階段的改善法可以套用於不同方法之後。文獻[6]中的 MHM 法即為一例，用來處理多變量資料集，是從文獻[15]的單變量 HM 變化而來，與不同方法(NPN、MD、MDAV、CBFS)結合，形成了 NPN-MHM、MD-MHM、MDAV-MHM、CBFS-MHM 四個方法。

三、 方法論

本研究提出之方法論主要是以傳統分群法中的「c-means」和「階層式分群」方法為基底，但由於原本 c-means 與階層式分群法的特性與不

[1] 以下文章均以 c-means 稱呼此方法，以避免與文中群內資料基數 k 意義混淆。

具群規模之限制條件的關係，使得我們直接使用這兩類方法來解微聚合問題時會遭遇到一些困難，例如：無法限制一群內資料點的數量在某一個範圍內，且事先給定群數也會限制住分群的可能性，使得分群結果品質降低。因此，我們提出兩個新的方法名為「KC」與「KCKW」，結合這兩類方法並作適當的修改，加入群規模的限制條件、讓群數在分群過程中動態的改變，以期找到一個最佳的分群結果。

(一) KC

此方法是以 c-means 分群法為基礎，加入群規模的限制條件(一群中至少包含 k 個、至多包含 2k-1 個資料元素)，並且讓群數動態的改變。演算法的 pseudo code 如圖 1 所示，說明如下：

A. 從資料集當中隨機挑選 c 個資料點當群心。
B. 計算每個資料點與 c 個群心的距離。
C. 將資料點輪流分群：針對每一個資料點，找到離其最近的一個群心，檢查此群心的群組內資料點數量，如果數量未達上限 2k-1，即加入此群心的群組；否則，不加入，繼續尋找下一個距離較近的群心，重覆前述檢查程序，直到將此資料點分派至某一群心的群組內為止。
D. 檢查每個群組內的資料數量，刪去數量為 0 的群組，針對數量未滿 k 個的群組 X 進行調整：計算在 X 中，k 減去群內資料數量的缺量，如果此缺量大於 X 內的資料數量，且在其他數量滿足 k 而未達 2k-1 的群組 Y 內仍有空缺大於或等於 X 的資料數量，那麼就將 X 內的所有資料點移至離各個資料點較近的群組 Y 中，接著刪去 X 這個群組；如果情況相反，則從所有群組內資料數量大於、等於k+1的群中選擇離X的群心最近的一筆資料到 X 裡，依次選取，直到補滿缺量為止。如此視情況調整群組，而不盲目補滿缺量的作法，可以適當去除掉不適合存在的群組，以防拉低整體的分群品質。
E. 移動群心：重新計算所有群組的群心，以每一群組內的資料取平均值成為新的群心。
F. 重覆步驟 B.到 E.，直到滿足所設定之停止條件為止，此處設定為滿足群心移動代數。

上述分群法得出的分群結果會是一個滿足群組規模限制(k~2k-1)的 k-分群。

(二) KCKW

在 Domingo-Ferrer 和 Mateo-Sanz 的研究[7]中，提出將階層式分群方法修改成具有群規模限制條件的分群法「k-Ward」，以符合微聚合問題的特性。而我們再將 KC 法與此文獻中的方法做一結合成為「KCKW」方法，希望去除掉 c-means 與階層式分群在微聚合問題中的限制後，還能夠結合兩者之優點，使得分群結果更為優秀。此演算法主要包含兩大部分，其 pseudo code 如圖 2 所示，說明如下：

一開始的作法和上述 KC 方法大致相同，但要修改步驟 C.與步驟 D.。將步驟 C.修改為：當資料點找到離其最近的一個群心後，即可加入此群心的群組內，不必再檢查此群組是否已達 2k-1 個資料數上限；而步驟 D.則改為：在調整群內資料數量未滿 k 個的群組時，如果其群組內缺量大於群組內的資料數量時，即可將群內所有資料點移至離各個資料點較近的其它群組中，不必再檢查其它群組中是否有空缺可容納。

```
KC (Dataset S, int c)
{
    P = φ;   //P is the partition result
    Select c points randomly in S to be centroids of c groups;

    While (!centroid moving iteration)
    {
        For (each x_i ∈ S)
        {
            Compute the distances between x_i and c centroids;
            Check the size of the closest group G_closest whose
            centroid is closest to x_i;
            While (!assign)
            {
                If (group size＜2k-1)
                {
                    Assign x_i to G_closest;
                  break;
                }
                else
                {
                    Check the size of the next closest group
                    G_closest whose centroid is closest to x_i;
                }
            }
        }
        For (each group G_i)
        {
            Check the group size of all groups;
            Remove the groups whose group size equals 0;
            If (the group size of G_i＜k)
            {
                //adjust the group size of G_i
                LackCount = k-| G_i |;
                Empty = Σ[(2k-1)-|other groups|];
                If (LackCount>| G_i | && Empty≧| G_i |)
                {
                    Assign each x_i ∈ G_i to the closest group
                    whose group size between k and (2k-2);
                    Remove G_i;
                }
                else
                {
                    For (1 to LackCount)
                    {
                        Move x_i from the closest group
                        whose group size is at least (k+1)
                        to G_i;
                    }
                }
            }
        }
        Compute the new centroid of each remaining group;
    }
    P = P ∪ all remaining groups;
    return P;
}
```

圖 1 KC 演算法 pseudo code

至此，分群的結果中，所有群組的群內資料數量都會滿足下限k，但不滿足上限2k-1的條件。因此，我們將群內資料數量超過2k-1的群組以k-Ward方法再次進行分群，其方法之pseudo code如下說明：

A. 以 MD 法在群組中找到最遠的兩個資料點，然後各自形成兩群，群組均為 k 的規模；剩下的點各自獨立一群。

B. 用華德法(Ward's)由下而上的聚合階層式分群，直到每一個資料點都屬於一個具有 k 筆或更多資料數量的群組。必須注意的是，在合併的過程中，不要把兩個同時具有 k 或者更多資料點的群組合併在一起。

C. 在目前的分群結果中，檢查是否有包含 2k 或更多資料數量的群組，如果有的話，再將這些群組以步驟 A.及步驟 B.的方式處理，直到所有群組都不再有超過 2k-1 個資料數量為止。

上述步驟執行完畢後，分群結果中的每一群組都會包含至少k個、至多2k-1個資料數量，也就是一個滿足群組規模限制(k~2k-1)的k-分群。

四、 實驗結果

本研究以微聚合問題的標竿資料集來進行實驗。以下簡介各個資料集，並將本研究之實驗數據與前人研究結果相比較。

(一) 資料集

本研究所採用之資料集是源自於歐洲的CASC(Computational Aspects of Statistical Confidentiality)計畫[1]，後續被許多此領域的研究者用來測試其方法論的好壞，各資料集規模及維度如下所述。

(1) Tarragona(834-13)：包含 834 筆資料，每筆資料具有 13 個數值型態的屬性。此資料集被使用於[6][7][17]的研究中。

```
KCKW (Dataset S, int c)
{
    P = φ;    //P is the partition result
    Select c points randomly in S to be centroids of c groups;
    While (!centroid moving iteration)
    {
        For (each x_i ∈ S)
        {
            Compute the distances between x_i and c centroids;
            Assign x_i to the group whose centroid is closest
            to x_i;
        }
        For (each group G_i)
        {
            Check the group size of all groups;
            Remove the groups whose group size equals 0;
            If (the group size of G_i < k)    //adjust the group
            {                                 //size of G_i
                LackCount = k-| G_i |;
                If (LackCount>| G_i |)
                {
                    Assign each x_i ∈ G_i to the closest group;
                    Remove G_i;
                }
                else
                {
                    For (1 to LackCount)
                    {
                        Move x_i from the closest group
                        whose group size is at least (k+1)
                        to G_i;
                    }
                }
            }
        }
        Compute the new centroid of each remaining group;
    }

    For (each group G_i)
    {
        Check the group size of G_i;
        If (the group size of G_i > (2k-1))
        {
            Use MD to form two groups with group size k;
            Let each remaining x_i ∈ G_i be a single group;
            While (!each x_i ∈ G_i is assigned to a group with
            group size more than k )
            {
                Compute the Ward's distances among groups;
                Combine the two closest groups as one group,
                but never combine both group sizes are more
                than 2k;
            }
        }
    }
    P = P ∪ all remaining groups;
    return P;
}
```

圖 2 KCKW 演算法 pseudo code

(2) Census(1080-13)：包含 1080 筆資料，每筆資料具有 13 個數值型態的屬性。此資料集被使用於[4][6][8][10][17][22][26]的研究中。

(3) EIA(4092-11)：包含 4092 筆資料，每筆資料具有 11 個數值型態的屬性。此資料集被使用於[6][22]的研究中。

(二) 實驗結果

本實驗設定安全參數 k 為 3，因此，在最後的分群結果中，每一群組必須包含 3 至 5 筆的資料，對於每個資料集均會測試 5 次，最後的數據結果取其平均值。在此將本研究之實驗數據(資訊損失量的百分比)繪製成圖 3-圖 5，並與前人研究中表現較好的幾個方法之數據互相比較以衡量方法之優劣。Uk-W(MD)之實驗數據出自文獻[7]，文獻中只採用 Tarragona 資料集來進行此方法的測試；M-d 之實驗數據出自於文獻[17]，其採用 Tarragona 及 Census 資料集進行實驗；NPN-MHM、MD、MD-MHM、MDAV、MDAV-MHM、CBFS、CBFS-MHM 實驗數據出自文獻[6]，在 EIA 資料集的實驗中，沒有針對 CBFS-MHM 方法進行測試。

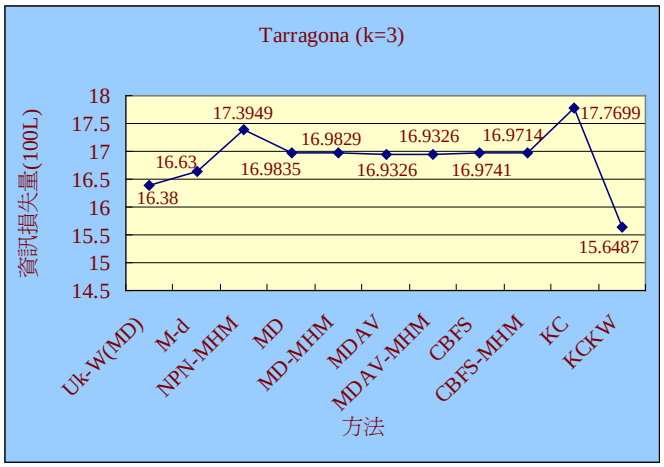

圖 3 Tarragona 資料集之實驗數據(k=3)

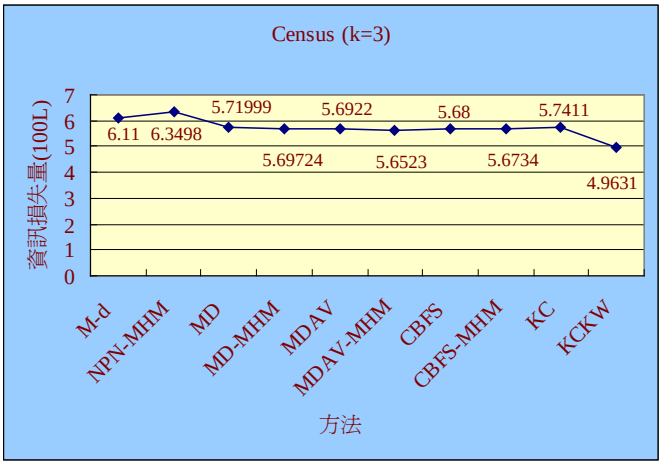

圖 4 Census 資料集之實驗數據(k=3)

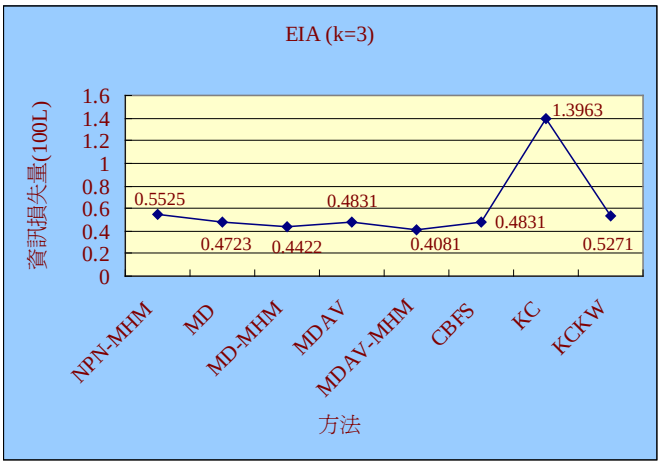

圖 5 EIA 資料集之實驗數據(k=3)

表 1 各方法針對各資料集所得之資訊損失量的表現排名與總和排名

Dataset / Method	Tarragona	Census	EIA	總和名次
Uk-W(MD)	2	N/A	N/A	N/A
M-d	3	9	N/A	N/A
NPN-MHM	9	10	6	25
MD	8	7	3	18
MD-MHM	7	6	2	15
MDAV	4	5	4	13
MDAV-MHM	4	2	1	**7**
CBFS	6	4	4	14
CBFS-MHM	5	3	N/A	N/A
KC	10	8	7	25
KCKW	1	1	5	**7**

觀察圖 3 及圖 4，我們可以發現在 Tarragona 和 Census 資料集的測試中，KC 法的表現不佳，但是 KCKW 法卻明顯的贏過所有方法，其資訊損失量最低。再者，就 Tarragona 資料集來看，KCKW 法勝過 KC 法和 Uk-W(MD)法，這說明了一件事：我們以 c-means 修改而成的微聚合方法透過與階層式分群法的結合，能夠大幅度的改善其效能，同時證實了我們一開始的想法「c-means 與階層式分群法的結合能夠改善各自的缺點並保留優點，比其中的任何一方表現更佳，發揮了一加一大於二的效果」。

然而，在圖 5 的折線圖中，KCKW 法的表現並不如前兩個資料集的測試結果般突出，只贏過了文獻中的一個方法 NPN-MHM。

接著我們對於各方法在各資料集的表現優劣作了整理，根據各自產生的資訊損失量將各方法排名(如表 1)，名次愈小代表表現愈好、資訊損失量愈小，顯示 N/A 的值即表示文獻中的此方法並未使用此資料集測試，因此無從得知。總和名次這一欄是由前三欄的名次加總而來，我們可以從這裡看出這些方法的綜合表現，其中 MDAV-MHM 方法和本研究提出的 KCKW 方法是表現最佳的兩個方法，若再將兩者就不同資料集的表現來比較，則各顯其優點，KCKW 在兩個資料集(Tarragona、Census)的實驗中均勝過了 MDAV-MHM 方法，而針對 EIA 的資料集來說，則是 MDAV-MHM 方法較為優秀。

五、 結論與未來發展

微聚合技術可以針對統計資料庫中的微資料作保護，使其具有匿名性，即使資料被發佈出去，也不會對個人隱私造成威脅，滿足 k-匿名的精神。然而，維護隱私的目標達成後，我們還需要兼顧資料的可用性，如此在後續的資料探勘工作中，才能確保探勘結果的正確性。這兩者是互相衝突的目標，若過度保護其中一方都會為另一方帶來負面的影響，因此，我們必須在這之間取

一適當的平衡點，同時保護個人隱私以及最小化資訊損失量以維持資料可用性。

過去有不少學者提出方法論來實現微聚合技術，其中以傳統分群法為基礎修改而成的方法論佔了不少比例，可見其重要性。據此，本研究也提出此類方法，並且進一步合併兩個分群法 c-means 及階層式分群法，融合彼此以截長補短，改善只以單一分群法為基礎修改的微聚合方法之效能，而後續的實驗結果也證實了我們的想法。至於往後的研究方向，或許能嘗試融合別種分群法，例如最小擴張樹的分群法，亦或是以與 GA 同為次經驗法則的其他種演化式計算(例如：粒子群最佳化方法、螞蟻演算法、禁制搜尋法及模擬退火煉鋼法等等)為方向，而這些方法都能夠再想想是否有互相結合的可能性，創造出單一方法無法達到的效能。

六、 致謝

本研究感謝國科會研究計畫的補助(計畫編號 NSC 98-2410-H-260-018-MY3)。

七、 參考文獻

[1] R. Brand, J. Domingo-Ferrer and J.M. Mateo-Sanz, “Reference data sets to test and compare sdc methods for protection of numerical microdata”, European Project IST-2000-25069 CASC, http://neon.vb.cbs.nl/casc, 2002.

[2] V. Ciriani, S. De Capitani di Vimercati, S. Foresti and P. Samarati, “k-anonymity”, Springer US, Advances in Information Security, 2007.

[3]]L.H. Cox, J.P. Kelly and R.J. Patil, “Computational Aspects of Controlled Tabular Adjustment: Algorithm and Analysis”, in: The Next Wave in Computer, Optimization and Decision Technologies (B. Golden, S. Raghavan and E. Wasil, eds.), Boston: Kluwer, 45-59, 2005.

[4] R. Dandekar, J. Domingo-Ferrer, F. Sebé, “LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection”, in: Inference Control in Statistical Databases (J. Domingo-Ferrer, ed.), vol. 2316 of LNCS, pp. 153–162. Springer, Berlin Heidelberg NewYork, 2002.

[5] D. Defays and P. Nanopoulos, “Panels of enterprises and confidentiality: the small aggregates method”, in: Proceedings of 1992 Symposium on Design and Analysis of Longitudinal Surveys, pp. 195–204, Statistics Canada, Ottawa, 1993.

[6] J. Domingo-Ferrer, A. Martínez-Ballesté, J.M. Mateo-Sanz and F. Sebé, “Efficient multivariate data-oriented microaggregation”, VLDB Journal, vol. 15, no. 4, pp. 355-369, 2006.

[7] J. Domingo-Ferrer and J.M. Mateo-Sanz, “Practical data-oriented microaggregation for statistical disclosure control”, IEEE Trans. Knowl. Data Eng. 14(1), 189–201, 2002.

[8] J. Domingo-Ferrer, J.M. Mateo-Sanz and V. Torra, “Comparing SDC methods for microdata on the basis of information loss and disclosure risk”, in: Pre-proceedings of ETK-NTTS’2001, vol. 2, pp. 807–826, Luxemburg, Eurostat, 2001.

[9] J. Domingo-Ferrer and V. Torra, “Fuzzy microaggregation for microdata protection”, Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 7, no. 2, pp. 153-159, 2003.

[10] J. Domingo-Ferrer and V. Torra, “Ordinal, continuous and heterogenerous k-anonymity

through microaggregation", Data Mining Knowl. Discov. 11(2), 195–212, 2005.

[11] F. Glover, L.H. Cox, R. Patil and J.P. Kelly, "Exact, Heuristic and Metaheuristic Methods for Confidentiality Protection by Controlled Tabular Adjustment", International Journal of Operations Research, vol. 5, No. 2, pp. 117-128, 2008.

[12] A.D. Gordon and J.T. Henderson, "An algorithm for Euclidean sum of squares classification", Biometrics, 33, 355–362, 1977.

[13] J. Han and M. Kamber, "Data mining: Concepts and Techniques", San Francisco: Morgan Kaufmann Publisher, 2001.

[14] P. Hansen, B. Jaumard and N. Mladenovic, "Minimum sum of squares clustering in a low dimensional space", J. Classifi. 15, 37–55, 1998.

[15] S.L. Hansen and S. Mukherjee, "Apolynomial algorithm for optimal univariate microaggregation", IEEE Trans. Knowl. Data Eng. 15(4), 1043–1044, 2003.

[16] A. Hundepool, A. Van de Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand and S. Giessing, "μ-ARGUS version 3.2 Software and User's Manual", Statistics Netherlands, Voorburg NL, http://neon.vb.cbs.nl/casc, 2003.

[17] M. Laszlo and S. Mukherjee, "Minimum spanning tree partitioning algorithm for microaggregation", IEEE Trans. Knowl. Data Eng. 17(7), 902–911, 2005.

[18] J.B. MacQueen, "Some methods for classification and analysis of multivariate observations", in: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297, 1967.

[19] J.M. Mateo-Sanz and J. Domingo-Ferrer, "A method for dataoriented multivariate microaggregation", in: Statistical Data Protection (J. Domingo-Ferrer, ed.), Luxemburg, Office for Official Publications of the European Communities, pp. 89–99, 1999.

[20] C.A. Murthy and N. Chowdhury, "In search of optimal cluaters using genetic algorithms", Pattern Recognition Letters, vol. 17, pp. 825-832, 1996.

[21] A. Oganian and J. Domingo-Ferrer, "On the complexity of optimalmicroaggregation for statistical disclosure control", Stat. J. United Nat. Econ. Com. Eur. 18(4), 345–354, 2001.

[22] A. Solanas, "Privacy Protection with Genetic Algorithms", Studies in Computational Intelligence (SCI), Springer, Berlin Heidelberg, 92, 215–237, 2008.

[23] V. Torra, "Microaggregation for categorical variables: a median based approach", in: Privacy Stat. Databases (J. Domingo-Ferrer and V. Torra, eds.), vol. 3050 of LNCS, pp.162–174, Springer, Berlin Heidelberg New York, 2004.

[24] J.H. Ward, "Hierarchical grouping to optimize an objective function", J. Am. Stat. Assoc. 58, 236–244, 1963.

[25] Rui Xu and D. Wunsch, II, "Survey of clustering algorithms", IEEE Transactions on neural networks, 16(3), 2005.

[26] W.E. Yancey, W.E. Winkler and R.H. Creecy, "Disclosure risk assessment in perturbative microdata protection", in: Inference Control in Statistical Databases (J. Domingo-Ferrer, ed.),

vol. 2316 of LNCS, pp.135–152, Springer,
Berlin Heidelberg New York, 2002.

A divisive partitioning method to microaggregation

Huei-Jhen Lin
National Chi Nan University
s97213519@ncnu.edu.tw

Peng-Yeng Yin
National Chi Nan University
pyyin@ncnu.edu.tw

ABSTRACT

The rapid development of information technology has made vast data collection easier. However, it also causes the concern about individual privacy that may be revealed from the data. To effectively protect the individual data stored in public statistical databases, the government can apply microaggregation which is one of statistical disclosure control techniques in addition to enforcing individual data protection laws. The problem nature of microaggregation is similar to that of classical clustering method. Consequently, there are several microaggregation techniques that are derived from classical clustering methods, disclosing the importance of this research direction. Based on this observation, we propose two new microaggregation methods, one is modified from c-means algorithm, and the other is created by marrying c-means and the hierarchical clustering methods. Our hybrid approach can remain the advantages of the two original methods without inheriting their drawbacks. Experimental results confirm our conjecture and manifest that our hybrid approach outperforms several existing methods.

Key Words: Microaggregation, Statistical disclosure control, Information loss, c-means clustering, Agglomerative approach