多次元クロス集計表における開示リスクと情報量損失の測定

独立行政法人統計センター　白川清美

1. はじめに

公的機関や学術研究に関わる統計結果表利用者から，既存結果表にない任意の多次元クロス集計が求められている。特に，新統計法施行以降，顕著に増えている。しかしながら，多様なニーズに対応した多次元クロス集計表を作成した場合，開示リスクの問題が発生する。さらに，このリスクを排除した場合，情報量の損失が伴う。また，これらの数値は測定できていない。そこで，本研究では，複数の既存結果表を組み合わせた確率分布に基づく多次元クロス集計表の作成を行う。さらに，この集計表における開示リスクと情報量損失の測定を行う。この研究の特徴は，既存結果表の確率分布に基づいた手法であるため，リンクによる開示リスクがなく，また，この計算工程を開示することにより一般利用者への提供が可能となる。さらに，ミクロデータから集計した真の集計値と比較することにより，情報量損失の測定が出来ることである。なお，先行研究として，ミクロアグリゲーションに関する研究動向及び匿名化技法として，超高次元クロス集計に基づいた質的属性の組合せパターンの検討[1]がある。ここでは，秘匿となる組合せの検討はしているが，新たなクロス表作成の検討はしていない。

2. 開示リスクと情報量の損失の評価

開示リスクは，結果表において「安全でないセル（度数が 1，2)」がある。そのリスクの分類は，「識別」，「属性開示（個人及び団体)」，「差分による開示」，と「開示の認識」がある。本研究における表開示抑制手法は，既存の公表済み結果表を複数組み合わせているため，集計後の手法を対象としている。多次元クロス集計では，各表の合計数に着目し，「安全でないセル」になる確率も求めている[2]。

情報量の損失は，データ供給者に対する手法に着眼している。特に，結果表における情報量の損失は，２次秘匿の数やセル値による開示リスクの分析や，平均情報量（エントロピー）がある [3] 。

確率変数 X のエントロピー：　$H(X) = -\sum_{k=1}^{n} p_k \log p_k$　$P(X = a_k) = P(a_k) = p_k, k = 1,2,\cdots,n$

さらに，各変数の有効性を評価するため，決定木の ID3 (Iterative Dichotomiser 3) [4]のエントロピーによる情報利得を算出している。この各変数におけるエントロピーを活用することにより，セルを秘匿することなく多次元クロス集計表の作成が出来ることと，計算方法を明示することにより開示リスクもなくなった。

3. まとめと今後の取組み

複数の結果表を組み合わせた多次元クロス集計表を作成した。また，集計に用いた計算手法を提示した。さらに，開示リスクと情報量の損失を評価した。その他，決定木の「よい識別の属性（エントロピーを下げる)」に基づいた深さが浅く接点数の少ないシンプルな ID3 のモデルが出来た。しかしながら，組み合わせ前の表に秘匿セルがある場合を考慮していない。それゆえ，今後の取組みは，集計前の結果表に秘匿処理がある場合におけるいくつかの暗号プロトコルと紛失多項式評価 (OPE: Oblivious Polynomial Evaluation) を用いた Cryptographic-Based Approach に基づく結果表作成である。

参考文献

[1] ミクロアグリゲーションに関する研究動向及び匿名化技法としてのミクロアグリゲーションの有効性に関する研究 ―全国消費実態調査を例に―，平成 20 年 9 月，独立行政法人統計センター ， 製表技術参考資料 10.

[2] Handbook on Statistical Disclosure Control, A Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control, 2010

[3] Measuring Disclosure Risk and Data Utility for Flexible Table Generators, Natalie Shlomo, Laszlo Antal and Mark Elliot, Work session on statistical data confidentiality, 2013

[4] Y. Lindel and B. Pinkas, Privacy Preserving Data Mining, Journal of Cryptology, Vol. 15, No.3, pp 177-206, 2002