Vol.2017-MPS-113 No.10
Vol.2017-BIO-50 No.10
2017/6/23

情報損失指標の非数値データへの適用

秋山 寛子[1,a)]　和田 昌昭[2]

概要：ミクロアグリゲーションにより失われる情報の量の割合を表す指標として，情報損失指標 ILD を定義する．ILD は，データ間の距離に基づき定義されるので，数値データだけでなく，適切な距離を選べば文字列で表現されるようなデータに対しても適用可能である．本論文では，ILD を非数値データセットに適用し，距離の設定について考察を行った．数値データセットに対するミクロアグリゲーションにおいてグループの代表値として平均値を用いる場合には，ILD は平均とデータの差の 2 乗和に基づいた情報損失指標と一致し，この情報損失指標を自然に拡張したものとなっている．

キーワード：情報損失，ミクロアグリゲーション，匿名化

Application of an Information Loss Index to Non-numerical Data

Hiroko Akiyama[1,a)]　Masaaki Wada[2]

Abstract: We define a new information loss index ILD as the ratio of the amount of information lost by microaggregation. Since ILD is based on distance between data, it is applicable not only to numerical data, but also to data like ones represented by character strings if we choose suitable distance. In this paper, we apply ILD to non-numerical datasets, and discuss choice of distance. For microaggregation of numerical datasets using the average values as the representatives of groups, ILD coincides with the information loss index based on the sum of squares of the difference between the average and data. In this sense, ILD is a natural extension of the information loss index.

Keywords: Information Loss, Microaggregation, Anonymization

1. はじめに

データセットをグループに分割し，グループごとにデータを代表値に置き換える操作はミクロアグリゲーションと呼ばれ，匿名化の一手法として研究されている（[1]，[2]）．あるデータセットにミクロアグリゲーションを実行すると，データセットの持つ情報の一部が失われる．このとき，失われる情報とはそもそも何なのか，一体どの程度の情報が失われたと考えれば良いのか，というようなことを考えたい．

本論文では，データ間の距離が任意に与えられたとき，それに基づいてデータセット全体の持つ情報の量を定義し，それを情報容量とよぶ．また，ミクロアグリゲーションにより失われる情報容量の割合として情報損失指標 ILD（Information Loss based on Distance）を定義する．

ILD は距離に基づき定義されており，数値属性以外の人名や住所のようなデータに対しても，適用することが可能となっている．3 節では，文字列で表されているデータセットに対してミクロアグリゲーションを実行し，様々なデータ間距離を設定して ILD を計算し，結果に対して考察を行う．

数値データセットに対するミクロアグリゲーションにおいては，分割されたグループごとの代表値としてグループの平均値を用いる場合がよく研究されており，平均との差の 2 乗和に基づいた情報損失指標 ILSSDM （Information Loss based on Sum of Square Difference from the Mean）

[1] 長野工業高等専門学校
Nagano Collage of Technology
[2] 大阪大学大学院情報科学研究科
Osaka University
[a)] h_akiyama@nagano-nct.ac.jp

Vol.2017-MPS-113 No.10
Vol.2017-BIO-50 No.10
2017/6/23

が用いられている（[3], [4], [5], [6], [7], [8], [9], [10]）．ILD は，ILSSDM を拡張したものであり，データ間の距離としてユークリッド距離を用いた場合，ILD はこの ILSSDM と同じ値になることを 4 節にて示す．

2. 情報損失指標 ILD の定義

2.1 情報容量の定義

データセット

$$A = (x_1, \ldots, x_N),\ \ x_i \in X (i = 1, \ldots, N)$$

があるとする．また，X の 2 つのデータ x, y 間の距離 $d(x, y)$ が与えられているとする．

X と距離については，例えば次のようなものを考えている．

- ユークリッド距離
  $X = \mathbb{R}^n$，$x, y \in \mathbb{R}^n$ に対し，
  $$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$
- 離散距離
  X は任意の集合，$x, y \in X$ に対し，
  $$d(x, y) = \begin{cases} 1 & (x \neq y) \\ 0 & (x = y) \end{cases}$$
- 木構造をもつデータセットの 2 つのデータ間距離
  X を任意の集合とする．データセット A（$A \subseteq X$）は，木構造を用いた階層型のミクロアグリゲーションが行われているとする．このとき，木の葉は A のデータであり，ノードは X のデータである．X の 2 つのデータ x, y 間の距離は，x から y へ到達するのに通る枝の数として求められる．

データセット A の情報容量 $I(A)$ を，

$$I(A) = \sum_{i=1}^{N} \sum_{j=1}^{N} d(x_i, x_j)^2$$

と定義する．

2.2 ILD の定義

データセット

$$A = (x_{11}, \ldots, x_{1n_1}, x_{21}, \ldots, x_{2n_2}, \ldots, x_{m1}, \ldots, x_{mn_m})$$

がグループ $A_1, \ldots, A_m$ に分割されているとする．ただし，

$$A_k = (x_{k1}, \ldots, x_{kn_k})\ \ (k = 1, \ldots, m)$$

である．いま，グループ A_k に属するデータをすべて A_k の代表値 $\hat{x}_k$ に置き換えると，データセットは，

$$\widehat{A} = (\hat{x}_1, \ldots, \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_2, \ldots, \hat{x}_m, \ldots, \hat{x}_m)$$

となる．このとき，一般に $I(A) > I(\widehat{A})$ となっている．

上記の置き換えに関する情報損失を，

$$ILD = \frac{I(A) - I(\widehat{A})}{I(A)}$$

と定義する．

以下に，ILD の具体例を示す．

例 2.1 (数値データセット) データセット $D = (1, 2, 3, 4)$ が $D_1 = (1, 2)$ と $D_2 = (3, 4)$ に分割されているとする．グループ D_1，D_2 に属するデータを，それぞれのグループの平均値に置き換えると，データセットは $\widehat{D} = (1.5, 1.5, 3.5, 3.5)$ となる．2 つのデータ間の距離をユークリッド距離で与えるとすると，D と $\widehat{D}$ の情報容量は，$I(D) = 40$，$I(\widehat{D}) = 32$ となる．したがって，平均値に置き換えることによる情報損失は，$ILD = \frac{40-32}{40} = 0.2$ である．

例 2.2 (記号データセット) 4 個の記号からなるデータセット $S = (a_{11}, a_{12}, a_{21}, a_{22})$ が，$S_1 = (a_{11}, a_{12})$ と $S_2 = (a_{21}, a_{22})$ に分割されているとする．S は，図 1 のように木構造をもつ階層型のミクロアグリゲーションがされているとする．グループ S_1, S_2 に属するデータを，それぞれ a_1，a_2 に置き換えると，データセットは $\widehat{S} = (a_1, a_1, a_2, a_2)$ となる．2 つのデータ間の距離をノード間の枝の数で与えると，S，$\widehat{S}$ の情報容量は，$I(S) = 144$，$I(\widehat{S}) = 32$ となる．したがって，情報損失は $ILD = \frac{144-32}{144} = 0.778$ である．また，S のデータをすべて a に置き換えると，データセットは $\widehat{\widehat{S}} = (a, a, a, a)$ となる．$\widehat{\widehat{S}}$ の情報容量は $I(\widehat{\widehat{S}}) = 0$ なので，情報損失は $ILD = \frac{144-0}{144} = 1$ である．

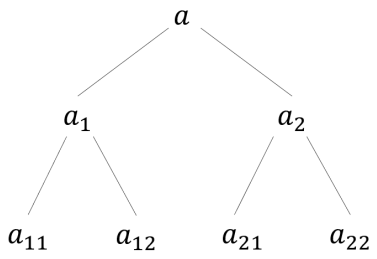

図 1 木構造をもつ記号データセット

3. ILD の非数値データへの適用

非数値データセットに対して，異なる距離を設定し ILD を計算し，考察を行う．

3.1 距離設定による ILD の比較

文字列で表される 8 個のデータからなるデータセット D について考える．都道府県のデータセット D=(長野，新潟，東京，神奈川，大阪，奈良，福岡，熊本) において，長野と新潟を甲信越に，東京と神奈川を関東に，大阪と奈良

Vol.2017-MPS-113 No.10
Vol.2017-BIO-50 No.10
2017/6/23

を関西に，福岡と熊本を九州に置き換えると，データセットは $\widehat{D}$=(甲信越，甲信越，関東，関東，関西，関西，九州，九州) となる．このように置き換えた場合，与える距離の設定によって ILD の値が変化する具体例を示す．

(1) 離散距離

データ間の距離を離散距離で与えると，D，$\widehat{D}$ の情報容量は，$I_{discr}(D) = 56$，$I_{discr}(\widehat{D}) = 48$ となる．したがって，$ILD_{discr} = 0.14285$ である．

(2) 木構造のデータ間距離

データセット D が図 2 のような木構造をもっている場合を考える．データ間の距離をノード間の枝の数で与えると，D，$\widehat{D}$ の情報容量は，$I_t(D) = 1440$，$I_t(\widehat{D}) = 576$ となる．したがって，$ILD_t = 0.6$ である．

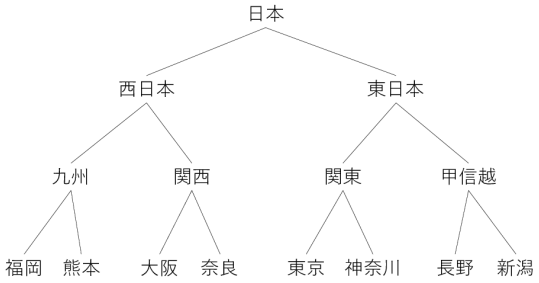

図 2 木構造をもつ都道府県データセット

(3) 重み付きグラフのデータ間距離

データセット D，$\widehat{D}$ が図 3，4 のような重み付きグラフの構造をもっている場合を考える．太い実線の重みは 6，太い破線の重みは 4，細い実線の重みは 2，細い破線の重みは 1 とする．データ間の距離を枝に付与された重みで与えると，$I_g(D) = 648$，$I_g(\widehat{D}) = 640$ となる．したがって，$ILD_g = 0.01234$ である．

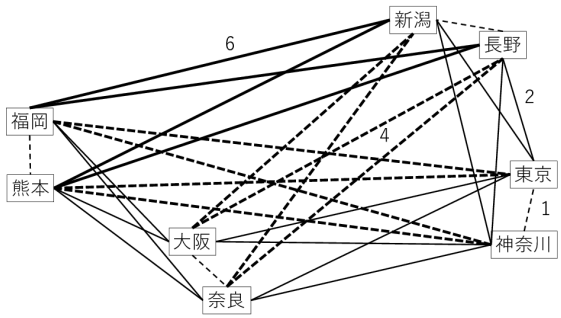

図 3 重み付きグラフの構造をもつ都道府県データセット D

以上の結果を表 1 に示す．

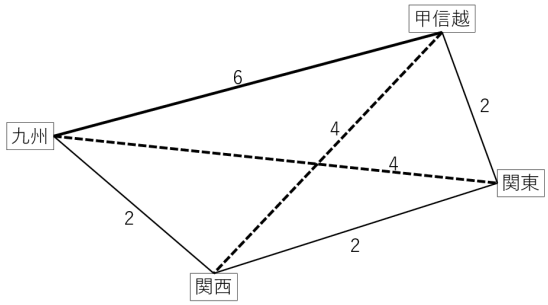

図 4 重み付きグラフの構造をもつ都道府県データセット $\widehat{D}$

表 1 異なる距離設定に対する情報容量と ILD の比較

	$I(D)$	$I(\widehat{D})$	ILD
離散距離	56	48	0.14285
木構造	1440	576	0.60000
重み付きグラフ	648	640	0.01234

3.2 考察

同一のデータセットに対して，異なる距離を設定して計算した ILD を比較し，情報容量の計算とデータ間の距離の設定という観点から考察する．

まず，情報容量の計算に関する考察を述べる．情報容量の定義にはデータ間の距離が含まれているが，$\mathbb{R}^n$ における距離としては，実数 $1 \leq p \leq \infty$ に対して定義される $p-$距離

$$d_p(x,y) = \begin{cases} (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} & 1 \leq p < \infty \\ \max_{i=1,\ldots,n} |x_i - y_i| & p = \infty \end{cases}$$

がある．$d_2(x,y)$ が 2.1 節にて定義した情報容量に用いられている距離である．3.1 節で計算した具体例について，$p = 1$ としたときの情報容量と ILD を表 2 に示す．どのような距離を用いるのが適切かは，応用しようとするデータの種類に応じた検討が必要である．

表 2 $p = 1$ としたときの情報容量と ILD の比較

	$I(D)$	$I(\widehat{D})$	ILD
離散距離	56	48	0.14285
木構造	272	160	0.41176
重み付きグラフ	168	160	0.04762

次に，距離の設定に関する考察を述べる．文字列などで表現されるデータや，カテゴリに分けられているデータなどに関して，データ種別に応じて距離の設定を考慮する必要がある．データのもつ意味や類似度などといったデータの特性を考慮した上で，データ間の距離を設定することにより，情報損失の度合いをより的確に把握できると考えられる．3.1 節で計算した具体例に関しては，データが異なっているかどうかのみを知りたい場合は離散距離，物理的な距離を考慮したい場合は重み付きグラフが，情報損失の把握に適していると考えることができる．

Vol.2017-MPS-113 No.10
Vol.2017-BIO-50 No.10
2017/6/23

4. ILD と ILSSDM の比較

数値データセットに対して，分割されたグループごとの代表値をグループの平均値とし，グループ内のデータを平均値に置き換えることを平均化とよぶ．本節では，数値データセットに対して平均化を行う場合の ILD が，ILSSDM と同じになることを示し，考察を述べる．

4.1 ILSSDM の定義

N 個のデータからなる数値データセット $X \in \mathbf{R}^n$ がグループ $X_1, \ldots, X_m$ に分割され，X_i が n_i 個のデータ x_{ij} $(0 \leq j \leq n_i)$ を含んでいるものとする．X_i の平均を $\overline{x}_i$ とし，平均からの距離の 2 乗和 (Sum of Squared Errors) を

$$SSE = \sum_{i=1}^{m} \sum_{j=1}^{n_i} |x_{ij} - \overline{x_i}|^2$$

で表す．一方，全データの平均を $\overline{x}$ とし，

$$SSA = \sum_{i=1}^{m} n_i |\overline{x_i} - \overline{x}|^2$$

とおけば，全データの平均からの距離の 2 乗和は

$$SST = \sum_{i=1}^{m} \sum_{j=1}^{n_i} |x_{ij} - \overline{x}|^2 = SSE + SSA$$

と表すことができる．このとき，ILSSDM は次のように定義される．

$$ILSSDM = \frac{SST - SSA}{SST} = \frac{SSE}{SST}$$

これは，X_i の分散 $var(X_i)$ と X の分散 $var(X)$ を用いれば，

$$ILSSDM = \frac{\sum_{i=1}^{m} n_i \, var(X_i)}{N \, var(X)} \tag{1}$$

と表すことができる．

4.2 情報容量の計算

3.1 節と同様の分割において，全データの情報容量は，

$$\begin{aligned} I(X) &= \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{x \in X_i} \sum_{y \in X_j} |x - y|^2 \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{x \in X_i} \sum_{y \in X_j} |(x - \overline{x}_i) + (\overline{x}_i - \overline{x}_j) + (\overline{x}_j - y)|^2 \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{x \in X_i} \sum_{y \in X_j} (|x - \overline{x}_i|^2 + |\overline{x}_i - \overline{x}_j|^2 + |\overline{x}_j - y|^2) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} n_i n_j (var(X_i) + |\overline{x}_i - \overline{x}_j|^2 + var(X_j)) \end{aligned} \tag{2}$$

となる．また，X の平均 $\overline{x}$ を用いた計算によって，

$$\begin{aligned} I(X) &= \sum_{x \in X} \sum_{y \in X} |x - y|^2 \\ &= \sum_{x \in X} \sum_{y \in X} |(x - \overline{x}) + (\overline{x} - y)|^2 \\ &= \sum_{x \in X} \sum_{y \in X} (|x - \overline{x}|^2 + |\overline{x} - y)|^2) \\ &= 2N^2 \, var(X) \end{aligned} \tag{3}$$

と表すこともできる．

4.3 平均化に関する ILD と ILSSDM の比較

X の分割 $X_1, \ldots, X_m$ において，各 X_i に含まれるデータをグループの平均値 $\overline{x}_i$ に置き換えたデータセットを X' とすると，

$$I(X') = \sum_{i=1}^{m} \sum_{j=1}^{m} n_i n_j |\overline{x}_i - \overline{x}_j|^2 \tag{4}$$

であるから，(2), (4) より，

$$\begin{aligned} I(X) - I(X') &= \sum_{i=1}^{m} \sum_{j=1}^{m} n_i n_j (var(X_i) + var(X_j)) \\ &= 2 \sum_{i=1}^{m} \sum_{j=1}^{m} n_i n_j \, var(X_i) \\ &= 2N \sum_{i=1}^{m} n_i \, var(X_i) \end{aligned} \tag{5}$$

したがって，(3), (5) より，

$$\begin{aligned} ILD &= \frac{I(X) - I(X')}{I(X)} \\ &= \frac{\sum_{i=1}^{m} n_i \, var(X_i)}{N \, var(X)} \end{aligned} \tag{6}$$

となり，ILD は ILSSDM に一致することがわかる．

4.4 考察

一般に，ILD の計算量は，データ総数を N とすると，$O(N^2)$ である．一方，ILSSDM の計算量は $O(N)$ であるから，$\mathbb{R}^n$ の数値データセットに対しユークリッド距離を用いる場合には，ILSSDM の計算式を用いることにより ILD を高速に計算することが可能である．

5. まとめ

本研究では，データセットにミクロアグリゲーションを実行する際に生じる情報の減少を客観的に示す指標として，情報損失指標 ILD を定義した．データセットのもつ情報の量として情報容量を定義することにより，元のデータセットに対して，ミクロアグリゲーションにより変化したデータセットはどの程度情報が減少したかを表すことが可能となった．ILD は，データ間の距離を用いて計算できるため，任意の距離が与えられれば，数値以外のデータに対

Vol.2017-MPS-113 No.10
Vol.2017-BIO-50 No.10
2017/6/23

しても情報損失を計算できる指標である．数値以外のデータとして都道府県のデータセットについて考え，ミクロアグリゲーションの結果に対して異なるデータ間距離の設定をし ILD を比較した．ILD 計算時のデータ間距離は，p–距離や離散距離など応用するデータの種類に応じて設定することができるため，情報損失の度合いをデータの特性に合わせてより的確に把握できると考えられる．また，ILD は ILSSDM を拡張したものとなっていることを，数値データセットに対して平均化を行う場合の ILD と ILSSDM が同じ値になることを証明することにより示した．

参考文献

[1] Domingo-Ferrer and Vicen c Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery, Vol. 11, No. 2, pp. 195–212, 2005.

[2] Domingo-Ferrer, Josep and Martínez-Ballesté, Antoni and Mateo-Sanz, Josep Maria and Sebé, Francesc, Efficient multivariate data-oriented microaggregation, The VLDB Journal—The International Journal on Very Large Data Bases, Vol. 15, No.4, pp. 355–369, 2006.

[3] Anthony WF Edwards and L Luka Cavalli-Sforza. A method for cluster analysis. Biometrics, pp. 362–375, 1965.

[4] AD Gordon and JT Henderson. An algorithm for euclidean sum of squares classification. Biometrics, pp. 355–362, 1977.

[5] Pierre Hansen, Brigitte Jaumard, and Nenad Mladenovic. Minimum sum of squares clustering in a low dimensional space. Journal of Classification, Vol. 15, No. 1, pp. 37–55, 1998.

[6] James MacQueen, et al. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, pp. 281–297. Oakland, CA, USA., 1967.

[7] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. Journal of the American statistical association, Vol. 58, No. 301, pp. 236–244, 1963.

[8] Agusti Solanas, Antoni Martinez-Balleste, and J Domingo-Ferrer. V-mdav: a multivariate microaggregation with variable group size. In 17th COMPSTAT Symposium of the IASC, Rome, pp. 917–925, 2006.

[9] Oganian, Anna and Domingo-Ferrer, Josep. On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the United Nations Economic Commission for Europe, Vol. 18, No. 4, pp. 345–353, 2001.

[10] Domingo-Ferrer, Josep and Mateo-Sanz, Josep Maria. Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and data Engineering, Vol. 14, No. 1, pp. 189–201, 2001.