匿名化による情報損失に関する研究

匿名化による情報損失に関する研究

秋山 寛子

平成 30 年 12 月

本論文は，奈良女子大学大学院人間文化研究科において
博士（情報科学）授与の要件として提出された博士論文である．

提出者：秋山 寛子

匿名化による情報損失に関する研究

秋山 寛子

概要

ビッグデータを活用し，新たな価値のある情報やサービスを生成することに関心が高まっている．ビッグデータには，気象情報のようなセンサデータや，防犯カメラの映像や SNS の画像データ，携帯端末の GPS による位置情報やクレジットカードの購買履歴といったパーソナル情報など，様々な種類の情報が含まれている．

パーソナル情報は，あらゆる場面で活用されている．位置情報に関しては，携帯端末の情報や設置されているセンサの情報などを利用し，防災計画や観光施策立案，交通情報の提供などに活用されている．購買履歴に関しては，企業のマーケティングや個人に対する商品のレコメンドなどに活用されている．医療情報に関しては，調剤履歴を利用した感染症流行状況の早期把握や，有害事象の早期発見などに活用できる可能性がある．ただし，個人に関連付けられる情報は，パーソナル情報であるため，その取り扱いには注意が必要である．

プライバシ保護の手法に，個人を特定・識別できないようにパーソナル情報を加工する匿名化という技術がある．匿名化により情報の匿名性を高くすればプライバシはより守られるが，一方で，情報はより減少してしまう．パーソナル情報の有効な活用のためには，匿名性と情報の有用性との最適なバランスが課題である．この課題について考える際には，それぞれの度合いを表す指標を用いる．匿名性を表す指標には，k-匿名性が広く用いられている．k-匿名性とは，データセットに含まれる任意のデータが，少なくとも他の $k-1$ 人と区別がつかない状態を表すものである．匿名化情報の有用性を測る指標としては，匿名化による情報の減少を測る情報損失指標がある．しかし，既存の情報損失指標は，特定のデータ種別や匿名化処理に対して定義されているものであり，その活用は限定的なものとなっている．そのため，より適用範囲の広い汎用性の高い情報損失指標が求められている．

本研究では，匿名化によりデータセットのデータの値が置き換えられたときに失われる情報を測る情報損失指標を考案する．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．データセットのデータの値を置き換える前後で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 の特徴は，データに対する適切な距離さえ与えれば情報損失が自動的に得られるという汎用性である．したがって，ILD を適用することでより多くのデータ種別に対して情報の損失を評価することが可能となる．

ILD は，異なるドメインやデータ種別の属性を含むデータセットに対しても，容易に適用が可能である．そのようなデータセットへの ILD の適用例として，アメリカの国勢調査のデータセットを用いて実験を行なった．k-匿名化の実現方法に，データセットを分割し同じグループ内のデータを同一の値に置き換えるミクロアグリゲーションという手法がある．データセットに対して異なるミクロアグリゲーションを行い，それぞれの ILD を計算し比較した結果，数値と非数値の両方の値に着目し，より近い値どうしを同じグループにしたものが最も ILD の値が小さくなった．ミクロアグリゲーションを行なったデータを活用する場合には，より似た値どうしが同じグループに含まれている方が，元のデータに対する情報の損失が少ないため活用に適していると言える．したがって，ILD は活用に適したデータセットの評価に有効であることを示した．

また，情報損失指標は匿名化アルゴリズムの開発への応用が可能である．例として，ミクロアグリゲーションにより k-匿名化されたデータセットに対して，分割を修正することにより情報損失を極小にするアルゴリズムの構成方法を示す．分割の修正を行うかどうかの判定条件は，情報損失指標の式を変形させることにより導かれた式であるため，理論的に必ず情報損失を極小にすることが可能である．

本論文の構成は次の通りである．第 1 章では，本研究の背景と目的を述べる．第 2 章では，パーソナル情報の匿名化について，データの種別や匿名化の手法，k-匿名性について述べる．また，k-匿名性をもたせる匿名化アルゴリズムや，匿名化により生じる情報の損失を測る指標について述べる．第 3 章では，提案する情報損失指標の定義を示し，具体的な計算例を示す．さらに，既存の情報損失指標と ILD の比較を行い，ILD の特徴について述べる．第 4 章では，ILD の適用と応用について述べる．数値と非数値の両方の属性を含む実データに対して，ILD を適用した結果を示し，これまでは難しかった数値と非数値の両方の類似度を反映した情報損失を表せることを示す．また，情報損失指標は匿名化アルゴリズムの開発へ応用することが可能であり，例として k-匿名性を保ちながら情報損失を極小にするアルゴリズムを構成する方法について述べる．第 5 章では，本研究の結論を述べる．

キーワード
匿名化，情報損失，ミクロアグリゲーション

A Study on Information Loss by Anonymization

Hiroko Akiyama

Abstract

There is growing interest in utilizing big data and generating new information and services. Big data includes various kinds of information such as sensor data for weather information, image of security cameras and images of SNS, personal information such as location information by GPS of mobile devices and purchase history of credit card, etc.

Personal information is utilized in every situation. Location information is utilized for disaster prevention planning, tourism planning, traffic information, etc. Purchase history data is used for marketing of companies and recommendation of products to individuals. Medical information has a possibility that it can be used for early detection of infection spread using dispensing history. However, since the information associated with individuals is personal information, we have to pay the care in dealing with them.

Anonymization is a method for privacy protection, it is a technique that processes personal information so that individuals cannot be identified or recognized. The more personal information has anonymity, the less the anonymized personal information has information. In order to utilize personal information effectively, we should consider the optimal balance between anonymity and usefulness of information. When considering it, we use indicators showing the degree of these. As an indicator for anonymity, k-anonymity is widely used. As an indicator for measuring the usefulness of anonymous information, there is an information loss indicator that measures reduction of information by anonymization. However, the existing information loss indicator is defined for a specific data type and anonymization processing, its use is limited. Therefore, the information loss indicator is required that is able to apply to many kinds of data type and anonymization processing.

In this research, we devise an information loss indicator that measures the information lost when the data value of the data set is replaced by anonymization. In this paper, we define Information Amount, which is the amount of information of

dataset based on the distance between data. We define the information loss indicator ILD (Information Loss based on Distance), which is the ratio of loss of the information amount by anonymization. ILD can be calculated automatically with the distance between data, and it is able to apply to many kind of datasets.

ILD can be easily applied to datasets including the variable kinds of attributes. As an example of application of ILD to such dataset, we applied it to American census dataset. Microaggregation is one of the methods for k-anonymization. We executed microaggregation algorithm to the dataset and applied ILD to the anonymized datasets. The smallest value of ILD is that of the anonymized dataset whose groups include the similar data with respect to both attributes. Therefore, we showed that ILD is effective for evaluating dataset for utilization.

Moreover, the information loss index is able to utilize to develop anonymization algorithms. As an example, we show a method of constructing an algorithm that minimizes information loss by modifying the division of k-anonymized dataset by microaggregation. The procedure is defined using the formula of information loss indicator, so we can guarantee that the algorithm make the information loss by anonimization minimum as long as using the procedure.

The organization of this thesis is as below. In chapter 1, we explain the background and the purpose of this research. In chapter 2, we explain the anonymization of personal information, types of personal information, the methods of anonymization, and k-anonymity. And also, we explain existing algorithms for k-anonymization and indicator for mesuring information loss by anonymization. In chapter 3, we define new indicator for measuring information loss and show some calculation examples of anonymized datasets. In chapter 4, we explain the applications of this information loss indicator. We show the results of application for a real dataset including both numeric and non-numerical data, Moreover, information loss indicator is useful for developing anonymization algorithm. For instance, we describe the way to making k-anonymization algorithm. In chapter 5, we make a conclusion of this thesis.

Keywords
anonymization, information loss, microaggregation

目 次

図 目 次

表 目 次

第1章　序論

ビッグデータを活用し，新たに価値のある情報やサービスを生成することに注目が集まっている．その背景には，スマートフォン等の様々なセンサデータを収集・流通できるデバイスの普及や，大規模データを処理する技術の発達などがある．大量のデータを収集・解析し，社会や個人に対する有益な情報の生成や研究への活用が期待されている．

ビッグデータにパーソナル情報が含まれている場合には，プライバシの保護が必要である．ビッグデータには，気象情報などのセンサデータや，防犯カメラの映像やSNSの画像データ，携帯端末のGPSによる位置情報やクレジットカードの購買履歴といった個人の活動履歴など，様々な種類の情報が含まれている．個人に関連付けられる情報は，パーソナル情報であるため，その取り扱いに注意が必要である．

パーソナル情報の活用が社会的に重要な問題となっているが，プライバシの保護と情報の有用性との最適なバランスが課題である．個人を特定・識別できないようにデータを加工する手法として匿名化がある．改正個人情報保護法では，匿名加工情報という定義が新たに設定され，匿名化されたパーソナル情報に対する需要はますます高まっている．匿名化を用いて匿名性を高めるとより安全なデータとなるが，一方で，情報の損失が生じるためデータの有用性が減少してしまう．

情報の活用に際しては，その有用性を測定し評価することが重要である．匿名化情報の有用性の尺度として情報損失指標がある．これは，匿名化により失われた情報を測定するものである．匿名化情報の生成に関する研究に注目が集まっているが，それらの多くが目標としているのは，匿名化による情報損失を小さくするアルゴリズム開発である．匿名化アルゴリズムの評価では，アルゴリズムにより生成された匿名化情報の情報損失を計測している．しかし，情報損失の計算式は論文によってそれぞれ異なっており，どのようなアルゴリズムに対しても，またデータセットに対しても適用できる統一的な指標が存在していないという問題がある．

本研究の目的は，匿名化によりデータセットのデータの値が置き換えられたときに失われる情報を測ることである．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．データセットのデータの値を置き換えたときに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 の特徴は，データに対する適切な距離さえ与えれば情報損失が自動的に得られるという汎用性である．ILD は，数値属性だけでなくカテゴリデータなどの非数値属性に対しても，数値と非数値の属性が混在するデータセットに対しても柔軟に適用できるため，より多くのデータセットに対して情報の損失を評価することが可能である．

本論文の構成は次の通りである．第 1 章では，本研究の背景と目的を述べた．第 2 章では，パーソナル情報の匿名化について，データの種別や匿名化の手法，k-匿名性について述べる．また，k-匿名性をもたせる匿名化アルゴリズムや，匿名化により生じる情報の損失を測る指標について述べる．第 3 章では，ILD の定義を示し具体的な計算例を示す．さらに，既存の情報損失指標と ILD の比較を行い，ILD の適用できるデータや応用先について述べる．第 4 章では，提案指標の適用と応用について述べる．数値と非数値の両方の属性を含む実データに対して ILD を適用し，これまでは難しかった数値と非数値の両方の類似度を反映した情報損失を表していることを示す．また，情報損失指標は匿名化アルゴリズムの開発へ応用することが可能であり，例として k-匿名性を保ちながら情報損失を極小にするアルゴリズムの構成に応用する方法について述べる．第 5 章では，本研究の結論を述べる．

第2章　匿名化と情報損失

パーソナル情報の活用に関心が高まっているが，その活用の際にはプライバシの保護が不可欠である．プライバシ保護の手法として，個人を特定・識別できないようにデータセットを加工する匿名化がある．匿名化を施すとプライバシを保護できる一方で，元のデータセットに比べて情報が減少してしまう．情報の有効活用のためには，匿名性と情報損失のバランスが課題である．匿名性については，k-匿名性が広く用いられ，k-匿名化についてよく研究されている．匿名化による情報損失を表す指標は，アルゴリズムやデータ種別によってそれぞれ定義されいる．

本章の構成は次の通りである．2.1 節では，パーソナル情報の定義や種類と，その匿名化の重要性について述べる．2.2 節では，匿名化の主な手法やその研究動向について説明する．2.3 節では，匿名化手法のミクロアグリゲーションについて説明する．2.4 節では，クラスタリングを用いたミクロアグリゲーションアルゴリズムについて詳細を説明する．

2.1　パーソナル情報の匿名化

2.1.1　パーソナル情報の定義

個人情報保護法では,『この法律において「個人情報」とは、生存する個人に関する情報であって、当該情報に含まれる氏名、生年月日その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ、それにより特定の個人を識別することができることとなるものを含む。）をいう。』とされている ([1])．また，OECD ガイドラインでは,『「個人データ」とは、識別された又は識別されうる個人（データ主体）に関するすべての情報を意味する。』とされている ([2])．

個人情報や個人データには，個人を特定できるような情報（名前や ID 等）だけでなく，個人の属性を表す情報（性別，所属等）や個人の活動履歴（位置情報，購買履歴等）も含まれている．後者に関しては，そのデータ単独では個人を特定でき

ない場合でも，複数の情報を突き合わせることにより個人を特定できる可能性がある．本論文では個人を特定できる情報と，個人を特定しうる情報とを合わせたものを「パーソナル情報」とし，その匿名化について考える．

2.1.2 属性値の種別

パーソナル情報とされるものには，画像，映像，音声，文章，個人の属性値などが含まれる．本論文では，個人の属性値の匿名化について考える．属性値とは，ある属性について個人がもつ値である．例えば,「出身地」という属性について，ある個人の属性値は「福岡」となる．

属性値は，数値データとカテゴリなどの非数値データに分類することができる ([3]). 以下にそれぞれの特徴を示す．

- 連続値 (continuous data)
  連続値は，数値属性の属性値である．このデータは，匿名化する際に算術的な操作を行うことができる．属性の例として，収入や年齢などがある．

- カテゴリデータ (categorical data)
  カテゴリデータは，カテゴリに分類されたデータである．このデータは，有限集合の値を取り，算術的な操作ができない．そのため，中央値などを計算したい場合は，新たな算出方法を設定する必要がある．カテゴリデータは，さらに順序データと文字データに分類される．

  - 順序データ (ordinal data)
    順序データは，順序関係をもつデータである．これらのデータには，データの比較（$\leq$）や最大，最小を求めるといった操作が可能である．属性の例として，命令レベルや優先順位などがある．
  - 文字データ (nominal data)
    文字データは，nominal 属性のデータであり，文字列などで表される．このデータは，順序を持たない．このデータへの可能なただ一つの操作は，同一かどうかの比較のみである．属性の例として，目の色や住所などがある．

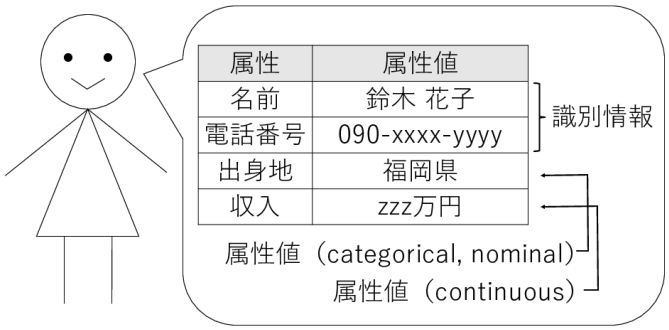

図 2.1: パーソナル情報の例

2.1.3 個人の識別，特定と匿名化

個人の「特定」とはある情報が誰の情報であるかわかることである．個人の「識別」とはある情報が誰か一人の情報であることがわかることである．匿名化情報とは，パーソナル情報から特定の個人の識別性を無くした情報である ([4])．これらの用語の定義を踏まえ，匿名化により生成される情報のカテゴリは次の 3 つとなる．

1. 識別特定情報：個人が識別されかつ特定される状態の情報
2. 識別非特定情報：一人に識別されるが，個人が特定されない状態の情報
3. 非識別非特定情報：一人に識別されずかつ個人が特定されない状態の情報

個人情報保護法における匿名化情報は，第二条 9 項に「匿名加工情報」として定義されており，個人識別符号の削除や他のデータへの置き換えなどにより,「特定の個人を識別することができないように個人情報を加工して得られる個人に関する情報」とされている（[5]）．本論文では，非識別非特定情報の生成方法と情報の有用性の評価について考える．

2.1.4 パーソナル情報の活用における匿名化の重要性

匿名化は，パーソナル情報を流通させ新たな価値のある情報やサービスを生み出すために必要な技術である．パーソナル情報の活用では，情報の提供者，情報の保持者，情報の分析者，情報の利用者という 4 つの立場が存在する ([6])．情報の提供

者は自身に関する情報を提供し，情報の保持者はそれを収集・蓄積する．情報の保持者は自組織でそれらの情報を解析できれば良いのだが，その技術がない場合には情報の分析者へデータを渡す．情報の分析者によって新たに価値のある情報が生成されると，それを用いて情報の利用者は有益なサービスを生み出す．情報の保持者は，情報の提供者に対しては収集した情報のプライバシを守る必要があり，一方で，データの分析者からはより情報量の多い有益な情報の提供をのぞまれている．

情報の保持者は，情報提供者の同意を得なければ第三者へのパーソナル情報の提供はできない．センサデータなどに関しては，個人が気づかない間にデータを収集されている場合もある．その場合，同意を取ることが困難なため，分析・活用するフェーズにまで個人に関する情報を到達させることができない．しかし，匿名化情報にした上で提供される情報の項目と提供方法を公表する場合，第三者への提供について同意が不要となる．したがって，匿名化により情報の分析者への提供が可能となるため，蓄積されたデータを活用し新たな価値を生み出すことができるようになる．

情報の保持者は，情報の分析者へパーソナル情報を提供する際には，個人を特定・識別できないように匿名加工処理をほどこす必要がある．データの一部の削除やあいまいなデータへの置き換えにより，データセットのもつ情報量は減少してしまうため，匿名化の方法によっては，情報の分析者は十分な分析や解析を行えないことがある．パーソナル情報の有効な活用のためには，分析者へ提供する情報の有用性がより高くなるような匿名化処理を考案する必要がある．

2.2 匿名化に関する研究

2.2.1 主な匿名化手法

パーソナル情報の属性値について，個人を特定・識別できないようなデータに加工する代表的な技術を以下に示す ([4],[7]).

- 属性情報の削除
  - 削除：個人を特定できる属性（氏名等）を削除する
  - 仮名化：個人を特定できる属性やその組み合わせを番号等に置換する
    （例：名前 → 社員番号）
- 属性情報の一般化
  - 一般化：属性の値をより一般的な値，あるいは広い値に置き換える．一般的には，階層木を生成する階層型一般化と，データ全体をあるルールに基づき分割しクラスタを生成する非階層型一般化がある (図 2.2).
    （例：奈良市 → 奈良県，25 歳 →20 代）
  - あいまい化：数値データについて，特に大きい（または小さい）値をまとめる
    （例：105 歳 →90 歳以上）
- 属性情報の加工
  - ノイズの付加：数値属性に対して，一定の分布に従う乱数的なノイズを加える
  - データ交換：レコード間で属性値を確率的に入れ替える
  - 擬似データ作成：元のデータと統計的に擬似している人工的な合成データを作成する
- その他
  - レコード削除:センシティブな属性値や外れ値などのレコードを削除する.
  - サンプリング:元データ全体から一定の割合・個数でランダムに抽出する

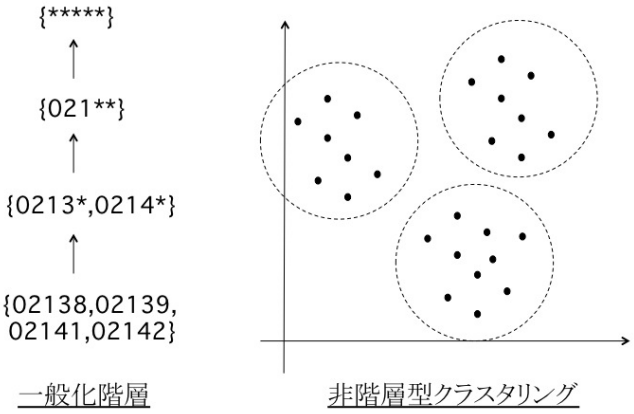

図 2.2: 階層型一般化と非階層型一般化

2.2.2 k-匿名性

パーソナル情報の匿名化にあたっては，どの程度匿名性を持たせるかを指標を用いて指定する必要がある．匿名性指標のうち，広く利用されているのが k-匿名性である ([8])．k-匿名性とは，データセットに含まれる任意の 1 つのデータが，少なくとも他の $k-1$ 個のデータと区別がつかない状態を表すものである．

k-匿名性を例を 2 つ示す．1 つ目の k-匿名化の例を図 2.3 に示す．7 つのデータを含む身長データセットを考える．左が元のデータセットで右が匿名化されたデータセットである．元のデータはそれぞれ異なる値なので，それぞれのデータは識別できる状態である．身長を 10cm で区切り値を置き換えると，150cm 台の人は 4 人と区別がつかなくなり，160cm 台の人は 3 人と区別がつかなくなる．匿名化されたデータセットは，どの任意のデータも少なくとも 3 人と区別がつかない．したがって，このデータセットは 3-匿名性をもつという．k-匿名性は，複数の属性を含むデータセットについても考えることができる．2 つ目の k-匿名化の例を図 2.4 に示す．住所と年齢の 2 つの属性を含むデータセットについて，住所は市町村から地方，年齢は 10 歳区切りの年代になるように匿名化する．各属性は 4-匿名性をもっているが，データセット全体としては 2-匿名性をもっている．

匿名性を表す指標は他にも，l-diversity([9])，t-closeness([10])，p-sensitive k-anonymity property([11])，Pk-anonymity([12]) などがあるが，いずれも k-匿名性を拡張したも

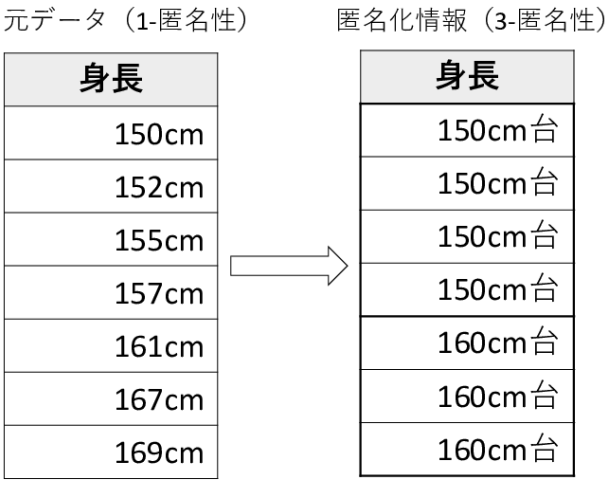

図 2.3: 1 つの属性をもつデータセットの k-匿名化の例

のである．k-匿名性は多くの指標の基本的な指標であるため，本論文では k-匿名性をもつ匿名化情報について考える．

k-匿名性が考案された背景として，匿名化情報の突き合わせによる再識別がある．あるデータセットがそれ単体では個人を特定できないように加工されていても，他のデータセットと突き合わせることで，個人を特定・識別されるリスクがある．

再識別の事例として，マサチューセッツ州の医療データと投票者名簿の突き合わせがある ([8])．医療データと投票者名簿に含まれていた属性を図 2.5 に示す．医療データと投票者名簿は，それぞれ個人を識別できる情報は削除された状態で公開されていた．それぞれのデータセットで共通の属性は，郵便番号，誕生日，性別である．マサチューセッツ州のウィリアム州知事は，マサチューセッツ州に住んでおり，彼のデータは医療データに含まれていた．投票者名簿によると，6 人は同じ誕生日であり，そのうち 3 人が男性であり，州知事の郵便番号をもつ人は一人しか存在しなかった．そのため，州知事の情報は一意に特定されてしまった．

このような再識別のリスクを回避するために，識別子の削除のみを施された匿名化情報を，複数人と区別できないようにデータを変更することを考える．そこで必要な考え方が k-匿名性である．k-匿名化により実現される匿名性とは，準識別子（個人を特定できる識別情報以外のデータ）に匿名性を与えることにより，外部にある他の情報と突き合わされても個人が特定されない状態である．

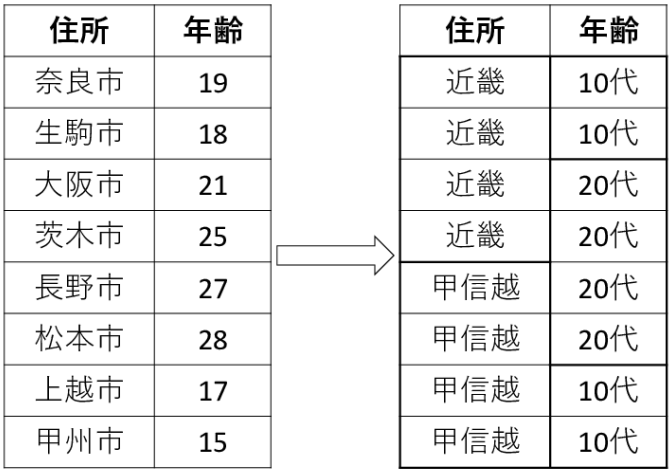

図 2.4: 複数の属性をもつデータセットの k-匿名化の例

2.2.3 匿名化の研究動向

匿名化は，一般化や一部のデータを秘匿にする方法を用いる手法が広く研究されている．一般化を用いて匿名化を行う方法には，データセットに含まれる属性のすべてのドメインを一般化する方法（[13]）や，一般化階層木の匿名化情報を自動生成する方法（[14]）などがある．また，一般化と一部のデータを秘匿にする方法を組み合わせた手法も研究されており（[15, 16]），匿名化によるデータの歪曲を抑えたり（[17, 18]），l-多様性へも対応できるよう拡張したり（[19]），処理速度を向上させたり（[20]）するなどの工夫がされている．

また，近似的なデータセットを作る方法も研究されており（[21, 22, 23, 24]），一部のデータを秘匿にして匿名化する手法において，秘匿にするデータを最小にする最適化について考察されている．

特定のデータ種別に対して応用可能な匿名化手法についても研究されている．具体的には，医療情報に関する応用（[25, 26]），テキスト情報に関する応用（[27]），位置情報に関する応用（[28, 29, 30]），大規模データに対する応用（[31, 32]）がある．

匿名化された情報から生成される統計情報に対するプライバシ保護に関する研究もされている．具体的には，収集したセンシングデータから統計情報を生成する際に，個人を識別されないようにする手法（[33]）や，データセットの分布が偏っているとき，データの分布を開示するとプライバシを保護できない場合，あいまい化を

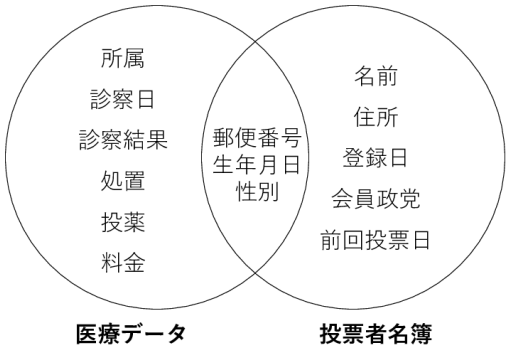

図 2.5: 匿名化情報の再識別の事例

用いてプライバシを保護する手法（[34]），撹乱を施された匿名化情報に対して統計情報を求める方法（[35]）などがある．

2.3 ミクロアグリゲーション

2.3.1 一般化と削除による k-匿名化の問題点

一般化と削除を用いて最適な k-匿名化を行うことは，NP 困難であることが知られている ([23]).

一般化を用いて k-匿名性を実現する場合，適当な匿名化データを設定するのにかかる実行時間が問題となる．カテゴリデータをもつ各属性について一般化した値を決定するには，各属性についてべき乗の数のパターンが存在する ([3]).

補助定理 1 c 個のカテゴリをもつ属性について，$2^c - c - 1$ の一般化の方法が存在する．

また，匿名化データとして新しいカテゴリを設定することと，その新しいカテゴリへ含める古いカテゴリを決定することも，最適な k-匿名化を求めるための手順が煩雑になる要因である．

また，削除を用いて k-匿名性を実現する場合，一般化と一部の削除の適切な組合せを決定する統一的な方法が存在しないという問題がある．さらに，削除とは値を空白にすることもしくは他の値に置き換えることにより実現されるため，分析などの際には，元のデータと同質の性質を持たなくなる場合があるという問題もある．

このように，一般化と削除の単一あるいは同時の使用には，k-匿名性をもたせるための統一的な方法がないため，データセットや秘匿にしたい属性やデータに依って，処理方法を変える必要がある．

2.3.2 ミクロアグリゲーションアルゴリズム

データセットに応じて一般化や削除の組合せや手順を考える必要がなく，統一した方法で処理できる k-匿名化手法がミクロアグリゲーションである ([36]). 最初はミクロアグリゲーションは連続値に対して定義されているものであったが ([37])，その後カテゴリデータへも拡張されている ([38]). ミクロアグリゲーションでは，どの種別のデータでも次の 2 つのステップにより匿名化することができる．

1. 分割 (partition) : 元データのうち似た値をもつ k 個以上のデータを含むようにグループを生成する．

2. 集約 (aggregation)：分割により生成されたグループに対して，集約値を計算する．例えば，連続値であれば平均値，カテゴリデータであれば中央値などとする．そして，元のデータをその集計値に置き換える．

分割された各グループ内のデータは，同じ値に置き換えられるため，識別できない状態となっている．一般化と削除を組合せた方法と比較して，ミクロアグリゲーションには次のような利点がある ([3])．

- ミクロアグリゲーションは，一般化と削除の組み合わせによる方法とは異なり，データ種別に依存しない統一された手順で実現できる．
- 適切なミクロアグリゲーションは NP 困難であるが，ほぼ最適なヒューリスティックスが存在する．
- 連続値をもつ属性については，数値的な意味を失うことなく適切にプライバシを保護することができる．

2.3.3 k 分割に関する研究

k 個のデータ以上から成るグループに分割する方法は，k-partition(k 分割) と呼ばれている（[37, 36]）．ミクロアグリゲーションによる k-匿名化を行うには，k 分割を行う必要がある．一般的なクラスタリングアルゴリズムでは 1 つのクラスタに k データ以上含まれることが保証されていないため，k 分割専用のアルゴリズムが研究されてきた（[39, 37, 40, 41, 36, 3, 42]）．

各グループに k 個以上のデータを含むようにするには，データ総数を N とすると N/k 個のグループに分割すればよいが，生成するクラスタの数を固定すると，データの密集している部分で分割したり，離れているデータどうしを同じグループにしてしまう可能性がある．たとえば，図 2.6 のようなデータセットを，k=2 としてグループの数を固定（N/k）して分割すると，図の左のように離れているデータが同じグループに含まれる可能性があり，似た値どうしが同じグループに含まれなくなってしまう．グループの数を固定しなければ，データの偏りを考慮した図の右のような分割にすることができ，より似た値どうしを含むグループを作ることができる．

最適な k 分割の方法は NP 困難であることが示されているため（[39, 43, 37]），計算量を抑えたアルゴリズムの開発や，より情報の歪曲や損失を抑えるようなヒューリスティックスの研究が行われている．後者の場合，各々情報損失を定義し，その値を抑えるように工夫したアルゴリズムが開発されている．

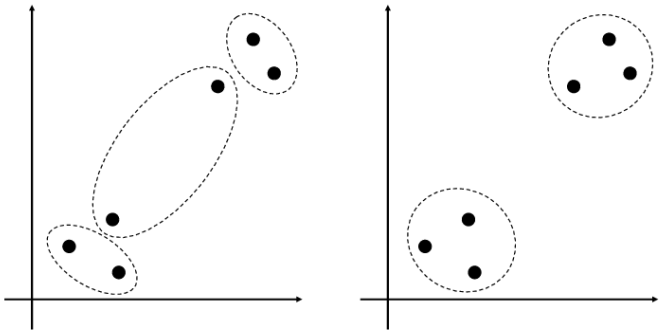

図 2.6: k 分割の例

k 分割を作成するヒューリスティックスは，空間分割に基づく手法（[44, 43]）と，クラスタリングに基づく方法（[39, 37, 40, 41, 36, 3, 42]）に大きく分けられる．空間分割に基づく手法では，データを多次元空間上の点と見なし，k 個以上のデータを含む空間を作れなくなるまで空間の分割をくり返す．分割には kd-tree（[45]）や R-tree（[46]）のような空間インデックスの構築アルゴリズムが使用される．クラスタリングに基づく手法には，木構造の深さに基づき分割する方法や，データ間の距離を用いて分割する方法がある．一般的なクラスタリングアルゴリズムでは 1 つのクラスタに k データ以上含まれることが保証できないため，k 分割専用のアルゴリズムが研究されてきた（[39, 37, 40, 41, 36, 3, 42]）．文献 [42] では，データの存在する距離空間においてクラスタの半径を最小にすることを目的とした r-gather r-celler アルゴリズムを提案している．文献 [39] では，クラスタに k 個以上のデータを含むように分割しつつ，クラスタ内の距離の総和を最小とする K-member clustering について，ランダムに初期値をとり非類似度の小さい k 個をクラスタとする Greedy K-member clustering を提案している．

2.3.4　クラスタリングを用いたミクロアグリゲーションアルゴリズム

本節では，ミクロアグリゲーションを用いたk-匿名化アルゴリズムであるMDAV([3])とVMDAV([36])の分割の部分についてその手順を示す．

アルゴリズムで使用する変数の定義は，次の通りである．

- R：データセット
- $|R|$：Rのデータ数
- D_i：分割されたRのi番目のクラスタ
- g：生成されたクラスタ数
- $\overline{x}$：Rの平均
- x_r：Rのうち$\overline{x}$から最も距離が大きいデータ
- x_s：Rのうちx_sから最も距離が大きいデータ
- c：Rの重心
- e_{min}：どのクラスタにも属しておらず，最も近いクラスタD_iのデータと最も近いデータ
- d_{in}：e_{min}と最も近いクラスタD_iのうち最も近いデータとの距離
- d_{in}：どのクラスタにも属しておらず，e_{min}と最も近いデータとの距離

MDAVアルゴリズム

MDAVは，アメリカの国勢調査や住宅調査などにも適用され実効性が確かめられている有名なアルゴリズムである．MDAVの分割方法の概要は，データセットの両端からk個ずつのクラスタを生成するアルゴリズムである．以下にMDAVアルゴリズムの内容を示す．

1. $|R| \geq 3k$である限り以下を繰り返す
   (a) $\overline{x}$を計算する
   (b) x_rを求める

(c) x_s を求める

(d) x_r, x_s の周りに 2 つのクラスタを作る．1 つのクラスタは，x_r と x_r と最も近い $k-1$ データを含む．もう 1 つのクラスタは，x_s と x_s と最も近い $k-1$ データを含む．

(e) R を 1.(d) にて生成したクラスタのデータを除いたデータセットに更新する．

2. $2k \leq |R| < 3k$ である限り以下を繰り返す

(a) $\overline{x}$ を計算する

(b) x_r を求める

(c) x_r と x_r と最も近い $k-1$ データを含むクラスタを生成する．

(d) どのクラスタにも属していない残りのデータで 1 つのクラスタを生成する．

3. $|R| < 2k$ の場合，どのクラスタにも属していないデータで 1 つのクラスタを生成する．

$k \leq l < 2k$ とすると，MDAV による分割の概要を図 2.7 に示す．特定の属性の種類

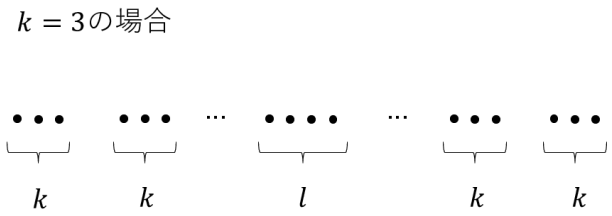

図 2.7: MDAV の分割の概略

についての実装には，平均値をどう計算するか，また，距離は何を使用するかを指定する必要がある．

V-MDAV アルゴリズム

VMDAV はデータの偏りを考慮し MDAV を改良したアルゴリズムである．VMDAV の分割方法の概要は，基本的には MDAV と同様に k 個のデータを含むクラスタを生成するものであるが，異なる点はクラスタの最も近傍の値をクラスタへ追加するかを判定する操作が追加されている点である．これにより，MDAV ではクラスタ内のデータ数が k に固定であるのに対し，VMDAV はクラスタ内のデータ数が可変となり，似た値どうしを 1 つのクラスタに含められるようになる．以下に VMDAV アルゴリズムの内容を示す．

1. すべてのデータの 2 点間の距離を計算する
2. c を計算する
3. どのクラスタにも属していないデータの数が k 以上である限り以下を繰り返す
   (a) c から最も遠いデータ x_r を求める
   (b) x_r と x_r と最も近い $k-1$ データを含むクラスタ D_i を生成し $R = R - D_i$ とする
   (c) e_min を求める
   (d) $d_{in} < \gamma d_{out}$ かつ $|D_i| < 2k-1$ である限り以下を繰り返す
      i. e_{min} を追加し $R = R - e_{min}$ とする
      ii. e_{min} を求める
4. どのクラスタにも属していないデータは，各データが最も近いクラスタへ追加する

VMDAV によるデータ追加の判定を図 2.8 に示す．

$\gamma = 0$ の場合，MDAV と同じアルゴリズムとなる．クラスタリングの操作で最も良い γ の値は，1 に近い値であるとされているが，最適な値については議論されていない．

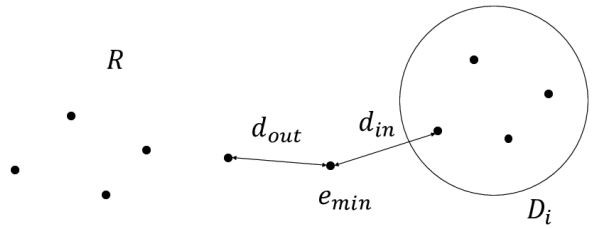

図 2.8: VMDAV の分割の概略

2.4 情報損失指標

2.4.1 匿名化情報の有用性の評価

本論文では，匿名化情報の有用性とは，その分析や活用に際してどれだけ有効活用できるかという観点から考えるものとする．匿名化情報を有効活用できる条件としては，次のようなものが挙げられる．

- 匿名化により秘匿にされるデータが少ない
- 元のデータセットのもつ統計的な量の変化が少ない
- 元のデータと匿名化された値との差が小さい
- 同じグループやカテゴリに分類された値どうしの類似度が高い

データセットは，匿名化処理をほどこすと，元のデータに比べて情報の一部を失う．削除という操作を行えば，データセットに含まれる一部情報がわからなくなり，さらにデータセット全体の統計的な値が変化する．一般化を行い階層木を作成する場合には，ノードの高さが大きくなるほど情報はあいまいなものとなり，元の情報との差異が大きくなる．ミクロアグリゲーションでは，同じグループに分類された値がグループの代表値に置き換えられるので，元のデータと代表値との差異が生じ情報が損失する．

匿名化により情報が失われることは，情報損失（Information Loss）とよばれている．匿名化処理に関する研究では，匿名化処理やデータセットの種類に応じて情報損失を定義し，情報損失が小さくなるような工夫がされている．つまり，情報損失の定義を活用することにより，匿名化情報の有用性を評価することが可能となる．

2.4.2 数値データセットに適用可能な情報損失指標

データセットの最適な分割とは，グループ内のデータどうしが似ていることである ([23], [37])．文献 [37] では，最適な k 分割は，グループ内の同質性を最大にすることとしており，文献 [47] では，グループの重心からの各データのユークリッド距離の総和が最小となることとしている．

分割の最適性を測る方法として，グループ内の各データ間とグループの重心とのユークリッド距離に基づいた値を使用している．この値は，匿名化の評価に限らず，

クラスタ分析において共通の尺度である ([48],[49],[50],[51],[52],[53]). データセットを k 個のクラスタへ分割する方法として，k-means 法があるが，設定される初期値によって分割結果が異なるという特徴がある．そこで，最適な分割を得るための手法を開発した MacQueen James らは，最適な分割を得る尺度として，グループにされた各点について，標本の確率分布を考慮した平均との誤差二乗和を使用している ([52]).

ミクロアグリゲーションでは，各グループに代表値を設定し，グループ内のデータを代表値へ置き換えるという操作を行う．数値データの場合，代表値は平均値や重心を用いるのが一般的である．本論文では，各データとグループの平均の距離に基づく情報損失を，ILSSDM（Information Loss based on Sum of Square Difference from the Mean）とよぶ.

ILSSDM の定義を示す．N 個のデータからなる数値データセット $X \in \mathbb{R}^n$ 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.

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

2.4.3 非数値データセットに適用可能な情報損失指標

匿名化による情報損失について，様々な算出方法が提案されている．本項では，k-匿名化アルゴリズムで用いられている，非数値データセットに適用可能な情報損失指標を示す.

Classification Metric

Classification Metric(CM)[54] は，クラス分けされた匿名化情報に対して適用する情報損失指標である．各データはクラスのラベルを付与されているものとする．CM は，各グループの多数をしめるクラス (マジョリティ) 以外のデータの数に基づき算出される．リレーショナルデータベースにおいて，テーブルの各行のペナルティ（マジョリティ以外のラベルがついているクラス）の合計を行の総数 N で割った値であり，次の式で定義される．

$$CM = \frac{\sum_{all\ rows} \mathrm{penalty}(row\ r)}{N}$$

数え上げられる行 r は，削除またはクラスのラベル $class(r)$ がグループ G のマジョリティのラベル $majority(G)$ でない場合にカウントされ，次のように表すことができる．

$$penalty(row\ r) = \begin{cases} 1 & \text{if } r \text{ is suppressed} \\ 1 & \text{if } class(r)\ \neq \mathrm{majority}(G(r)) \\ 0 & \text{otherwise} \end{cases}$$

例として，図 2.9 の場合を考える．属性は X と Y の 2 つであり，データセットに含まれる 20 個のデータは 4 つのグループ $G_1, \ldots, G_4$ に分割されている．各データは 2 つのクラスが付与されており，それぞれ ◯ または △ で表している．G_1 はすべて同じクラスのデータであるが，G_2, G_3, G_4 はそれぞれ，1 つずつマジョリティでないデータがあるため，データセット全体のペナルティは 3 となる．したがって，$CM = \frac{3}{20}$ となる．

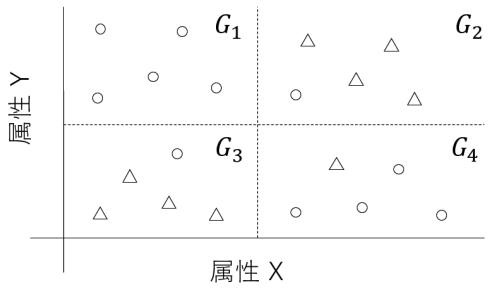

図 2.9: CM の計算の例

Discernibility Metric

Discernibility Metric(DM)[55] は，一般化と削除を用いた匿名化に対する情報損失指標である．DM は，匿名化されたデータセットのどれだけのタプルが識別不可能かに基づいて算出する．DM を用いて評価する匿名化アルゴリズムでは，k 値を設定し，グループのデータ数が k 未満のものは認識されないとしている．D をデータセット，$|D|$ を D のデータ数，E を匿名化処理 g によりグループ分けされた D のうち同じグループのデータの集合とすると，DM は次の式で求められる．

$$C_{DM}(g,k) = \sum_{\forall E_{s,t}|E|\geq k} |E|^2 + \sum_{\forall E_{s,t}|E|<k} |D||E|$$

例として，$|D| = 8, k = 3$ の場合を考える．4 つと 4 つのデータを含む 2 つのグループに分けられた場合，$C_{DM} = 4^2 + 4^2 = 32$ となり，1 つと 3 つと 4 つのデータを含む 3 つのグループに分けられた場合，$C_{DM} = 3^2 + 4^2 + 1 \cdot 8 = 33$ となる．

Normalized Certainty Penalty

Normalized Certainty Penalty(NCP)[56] は，匿名化による情報損失と属性の重要性という 2 つの観点に基づき算出する情報損失指標である．

数値データを持つ属性について考える．T を準識別子 $(A_1, \ldots, A_n)$ を列とするテーブルとする．タプル $t = (x_1, \ldots, x_n)$ を一般化し，$t' = ([y_1, z_1], \ldots, [y_n, z_n])$ とする．ただし，$y_i \leq x_i \leq z_i (1 \leq i \leq n)$ とする．属性 A_i について，NCP は次の式で定義される．

$$NCP_{A_i}(t) = \frac{z_i - y_i}{|A_i|}$$

ただし，$|A_i| = \max_{t \in T}\{t.A_i\} - \min_{t \in T}\{t.A_i\}$ は属性 A_i におけるすべてのタプルの範囲である．

各属性 A_i に対して匿名化情報の分析における有用性に基づき重み w_i を持たせるとすると，タプル t について，Weighted Certainty Penalty は以下の式で求められる．

$$NCP(t) = \sum_{i=1}^{n} (w_i \cdot NCP_{A_i}(t)) \tag{2.1}$$

$$= \sum_{i=1}^{n} (w_i \cdot \frac{z_i - y_i}{|A_i|}) \tag{2.2}$$

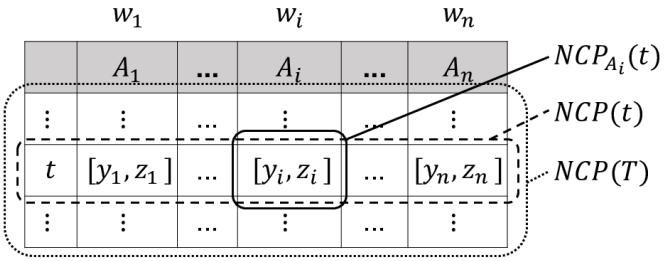

図 2.10: NCP の計算における各変数の対象となる範囲

このようにして，全ての属性とタプルについて算出した値の総和が，有用性に基づいた指標であり，$NCP(T) = \sum_{t \in T} NCP(t)$ で求められる．

次に，カテゴリデータをもつ属性について考える．カテゴリデータを階層木をもつように匿名化する場合について考える．例えば，郵便番号は地域，市，などのように階層木で表せるような構造になっている．$v_1, \ldots, v_l$ を階層木の葉とする．u を節とし，$v_1, \ldots, v_l$ の共通祖先とする．ただし，u は $v_1, \ldots, v_l$ の祖先となるような子孫はもたない．u の子孫の葉の数は $\mathrm{size}(u)$ とよぶ．このとき，属性 A に対するタプル t の NCP は次の指揮で定義される．

$$NCP_A(t) = \frac{\mathrm{size}(u)}{|A|}$$

ただし，$|A|$ は属性 A のもつ値の数である．

最後に，数値とカテゴリデータを両方持つデータセットについて考える．NCP は次の式で定義される．

$$NCP(T) = \sum_{t \in T} \sum_{i=1}^{n} (w_i \times NCP_{A_i}(t))$$

$NCP_{A_i}(t)$ は数値データ，カテゴリデータそれぞれに定義された式を用いて計算する．

第3章　情報損失指標ILD

匿名化アルゴリズムの評価には情報損失指標が用いられているが，既存の情報損失指標は特定の匿名化手法やデータ種別にのみ適用可能なものとなっており，広く活用できる性質をもっていないという問題がある．本論文では，データ間の距離さえ与えれば情報損失を計算できる，汎用性の高い情報損失指標ILD（Information Loss based on Distance）を提案する．

本章の構成は次の通りである．3.1節では，本論文で提案する情報損失指標ILDの定義を示す．3.2節では，非数値データのデータ間距離の設定方法やILDの計算例を示す．3.3節では，既存の情報損失指標とILDを比較し，ILDのもつ特徴について述べる．

3.1　情報損失指標ILDの定義

3.1.1　情報損失指標ILDの概要

匿名化アルゴリズムの開発においては，その効果を情報損失指標を用いて評価している．評価に用いる情報損失指標は，提案されるアルゴリズムごとに異なるものが使われており，どのようなアルゴリズムやデータ種別に対しても適用できる統一的なものが存在していない．

提案する情報損失指標には，次の3つの性質を持たせることを目標とする．1つ目は，元のデータセットと比較した情報損失を表していることである．2つ目は，データの類似度を反映した情報損失指標となっていることである．3つ目は，あらゆるデータ種別に対して適用可能であることである．

1つ目の性質に関しては，データセットそのもののもつ情報の量を定義し，データセットに含まれるデータの置き換えの前後で，情報の量の差により情報損失指標を定義する．2つ目の性質に関しては，データセットそのもののもつ情報の量を，データ間の距離を用いて定義することにより実現する．さらに，非数値データについて

は，似ているデータどうしの距離を小さくし，似ていないデータどうしの距離を大きくすることで，類似度を表すことが可能になる．3 つ目の性質に関しては，数値と非数値の両方のデータを含むデータセットへも適用可能なものとする．

3.1.2 情報容量と ILD の定義

まず，データセットそのものが持つ情報の量として，情報容量（Information Amount）を定義する．$I(A)$ は，データどうしの関連性をデータ間の距離によって与えた上でデータセット A に含まれる情報の量を測るものである．

データセット

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

があり，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．このとき，データセット A の情報容量 $I(A)$ を，

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

で定義する．なお，情報エントロピーと区別するために，情報量とはよばず情報容量とよんでいる．

次に，ILD の定義を示す．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．

3.1.3　データ間の距離と計算例

情報容量の計算に用いる距離 $d(x,y)$ については，

(i) $d(x,y) \geq 0$　かつ　$d(x,y) = 0 \Leftrightarrow x = y$

(ii) $d(x,y) = d(y,x)$

を仮定するが，三角不等式は成り立たなくても以下の議論に影響はない．具体的には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，

$$d(x,y) = \begin{cases} 1 & (x \neq y) \\ 0 & (x = y) \end{cases}$$

- 直積距離
  d_1, d_2 をそれぞれ X_1, X_2 上の距離とするとき直積集合 $X = X_1 \times X_2$ 上の距離が

$$d((x_1,x_2),(y_1,y_2)) = \sqrt{w_1 d_1(x_1,y_1)^2 + w_2 d_2(x_2,y_2)^2}$$

  で定義される．重み w_1, w_2 は任意の正の実数である．

- 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．

以下に，ILD の具体的な計算例を示す．

例 3.1.1 (数値データセット) データセット $A = (1, 2, 3, 4)$ が $A_1 = (1, 2)$ と $A_2 = (3, 4)$ に分割されているとする．グループ A_1，A_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 $\widehat{A} = (1.5, 1.5, 3.5, 3.5)$ となる．データ間の距離をユークリッド距離で与えるとすると，A と $\widehat{A}$ の情報容量はそれぞれ，$I(A) = 40$，$I(\widehat{A}) = 32$ 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 3.1.2 (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 *3.1* 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．

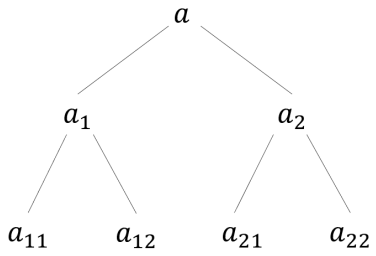

図 3.1: 木構造をもつ記号データセットの例

例 3.1.3 (数値と記号の組) 数値と記号の組からなるデータセット $A = ((1, a), (2, a), (3, b), (4, c))$ が $A_1 = ((1, a), (2, a))$ と $A_2 = ((3, b), (4, c))$ に分割されているとする．グループ A_1, A_2 に属するデータをそれぞれ $(1.5, a)$ と $(3.5, b)$ に置き換えると，データセットは $\widehat{A} = ((1.5, a), (1.5, a), (3.5, b), (3.5, b))$ となる．数値の距離をユークリッド距離，記号の距離を $d(a, b) = 1$，$d(b, c) = 1$，$d(a, c) = 3$ とし，直積距離を用いてデータ間の距離を計る．数値データと記号データを対等に評価するために重みをそれぞれ数値データのみの情報容量，記号のみの情報容量の逆数として $w_1 = 1/40$, $w_2 = 1/42$

とすると，$I(A) = 2,\ I(\widehat{A}) = 104/105$ となる．したがって，この場合の情報損失は $ILD = \frac{2-104/105}{2} = \frac{53}{105}$ である．

3.2　データ間距離の設定

3.2.1　非数値データの距離の設定例

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．

非数値データセットに対して，同一のデータの置き換えをする場合に，異なる距離を設定しILDを計算し値を比較する．例として，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3.2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．

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3.3，3.4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．

以上の結果を表3.1に示す．

本項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．

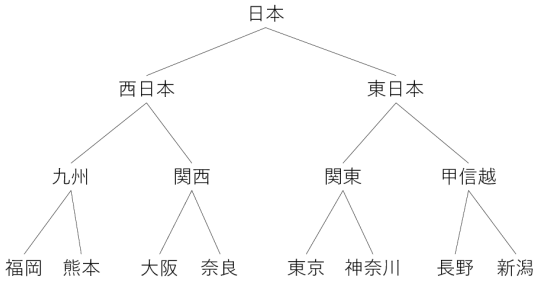

図 3.2: 木構造をもつ都道府県データセット

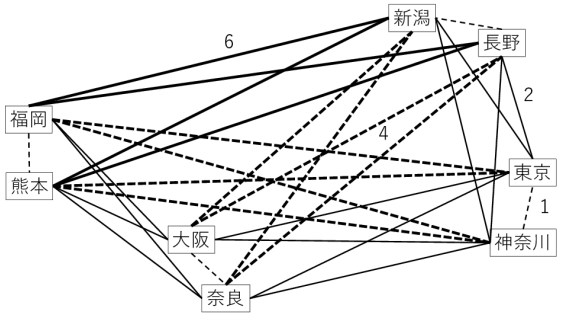

図 3.3: 重み付きグラフの構造をもつ都道府県データセット D

3.2.2　文字列データの距離

文字列で表されるデータ間の距離は，文字列の類似度で考えることができる．文字列データの距離を類似度で与える場合，個々の距離を個別に設定することなく，計算により自動的に得られるという利点がある．文字列の類似度を測る方法として，次のようなものがある．

ハミング距離

ハミング距離とは，長さが等しい 2 つの文字列の類似度を測るものであり，違う要素の数を数え上げるものである．長さ n の文字列 $x = x_1, \ldots, x_n$，$y = y_1, \ldots, y_n$

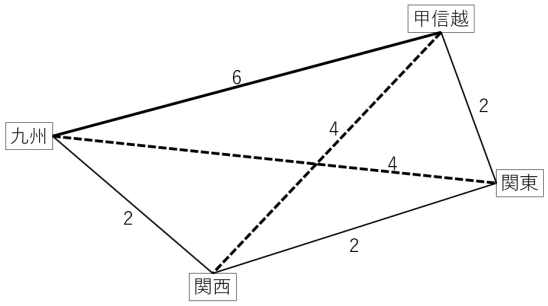

図 3.4: 重み付きグラフの構造をもつ都道府県データセット $\widehat{D}$

表 3.1: 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

のハミング距離 $d_H(x, y)$ は，次の式で定義される．

$$d_H(x, y) = \sum_{i=1}^{n} d_D(x_i, y_i)$$

ただし，$d_D(x_i, y_i)$ は x_i，y_i の離散距離である．

最小編集距離

最小編集距離とは，ある文字列をどれだけ編集すれば別の文字列になるかに基づき計算される距離である．編集とは,「挿入」,「削除」,「置換」の 3 つを表す．挿入とは，文字列に文字を 1 つ入れる操作である．削除とは，文字列から文字を 1 つ消す操作である．置換とは，文字列の中の 1 つの文字を異なる 1 文字に置き換える操作である．

レーベンシュタイン距離

レーベンシュタイン距離とは，最小編集距離の一種で，挿入，削除，置換が行われる回数に基づき計算される距離である．一般的に文字列が短いほど距離は小さく，文字列が長いほど距離が大きくなるため，文字列の長さで割った，標準化されたレーベンシュタイン距離（Normalized Levenshtein Distance）が広く用いられている．

Damerau-Levenshtein 距離

Damerau-Levenshtein 距離とは，レーベンシュタイン距離の 3 つの編集操作に「転置」を加えて計算される距離である．転置とは，隣接する 2 つの文字の位置を入れ換える操作である．

3.2.3 p-norm の p 値の設定

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 $p-$ 距離があり，次のように表される．

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

$d_2(x,y)$ が 3.1 節にて定義した情報容量に用いられている距離である．

$p = 1$ はマンハッタン距離，$p = 2$ はユークリッド距離，$p = \inf$ はチェビシェフ距離である．チェビシェフ距離は，複数の属性を持つデータの場合，1 つの属性値でも大きく異なると全く異なるデータであると判断したい場合に有用である．

前項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 3.2 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 3.2: $p = 1$ 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

3.3 既存の情報損失指標と比較したILDの特徴

3.3.1 情報損失指標ILSSDMとの比較

2.2.2 項で示した，数値データセットに適用可能な情報損失指標 ILSSDM と ILD の比較を行う．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.

N 個のデータからなる数値データセット $X \in \mathbb{R}^n$ 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 $\overline{x_i}$，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，全データの平均からの距離の 2 乗和を

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

と表す．このとき，ILSSDM は次のように定義される．

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

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

$$ILSSDM = \frac{\sum_{i=1}^{m} n_i \, var(X_i)}{N \, var(X)} \tag{3.1}$$

と表すことができる．

次に，同様の分割における情報容量を示す．同様の分割において，全データの情

報容量は，

$$
\begin{aligned}
I(X) &= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{x\in X_i}\sum_{y\in X_j}|x-y|^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{x\in X_i}\sum_{y\in X_j}|(x-\overline{x_i})+(\overline{x_i}-\overline{x_j})+(\overline{x_j}-y)|^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{x\in X_i}\sum_{y\in X_j}(|x-\overline{x_i}|^2+|\overline{x_i}-\overline{x_j}|^2+|\overline{x_j}-y|^2) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} n_i n_j (var(x_i)+|\overline{x_i}-\overline{x_j}|^2+var(x_j)) \qquad (3.2)
\end{aligned}
$$

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

$$
\begin{aligned}
I(X) &= \sum_{x\in X}\sum_{y\in X}|x-y|^2 \\
&= \sum_{x\in X}\sum_{y\in X}|(x-\overline{x})+(\overline{x}-y)|^2 \\
&= \sum_{x\in X}\sum_{y\in X}(|x-\overline{x}|^2+|\overline{x}-y|^2) \\
&= 2N^2\, var(X) \qquad (3.3)
\end{aligned}
$$

と表すこともできる．

最後に，平均化に関する ILD と ILSSDM を比較する．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，

$$
I(X') = \sum_{i=1}^{m}\sum_{j=1}^{m} n_i n_j |\overline{x}_i-\overline{x}_j|^2 \qquad (3.4)
$$

であるから，(3.2)，(3.4) より，

$$\begin{aligned} I(X) - I(X') &= \sum_{i=1}^{m}\sum_{j=1}^{m} n_i n_j (var(X_i) + var(X_j)) \\ &= 2\sum_{i=1}^{m}\sum_{j=1}^{m} n_i n_j \, var(X_i) \\ &= 2N \sum_{i=1}^{m} n_i \, var(X_i) \end{aligned} \tag{3.5}$$

したがって，(3.3)，(3.5) より，

$$\begin{aligned} ILD &= \frac{I(X) - I(X')}{I(X)} \\ &= \frac{\sum_{i=1}^{m} n_i \, var(X_i)}{N \, var(X)} \end{aligned} \tag{3.6}$$

となり，(3.1)，(3.6) より，ILD は ILSSDM に一致することがわかる．したがって，ILD は ILSSDM の拡張となっている．

ILSSDM と比較した場合の ILD の特徴は，ユークリッド空間でなくても計算が可能である点である．数値属性と非数値属性を含むデータセットについて，データ間距離として直積距離を設定した場合，非数値データセットのデータ間距離をユークリッド距離に埋め込めるとき，ILSSDM と ILD は同じ値になる．しかし，2.1 節 例 2.2 のようにユークリッド空間に埋め込めない距離をもつデータセットも存在し，その場合は ILSSDM によって情報損失を評価することができない．

計算量に関して，ILSSDM と比較し考察を述べる．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．

3.3.2 非数値データセットに適用可能な情報損失指標との比較

非数値データセットに適用可能な情報損失指標は，データの個数に基づいて計算するものや，一般化の場合データのとる値の範囲に基づくものが多い．データの個数という情報のみからであると，データセットからどのようなデータがどういうデータへ置き換わったのかという，データどうしの類似度に関する情報が反映されてい

ない．また，データのとる値の範囲に基づくものについては，元のデータセットの個々のデータが，置き換えられたデータ範囲の中でどのように分布しているかの情報は反映されていない．

Classification Metric(CM)[54] は，複数の属性を持つデータセットへも適用可能であるが，カテゴリデータの組に対してラベルが付与されるため，連続値をもつ数値属性とカテゴリ属性が両方含まれるデータセットには適用できない．

Discernibility Metric（DM）[55] は，グループに含まれるデータ数を用いて計算される．例えば，異なる 5 つの値のデータから成るデータセットを，データ数が 2 つと 3 つのグループに分割する場合，どのデータを同じグループにしても DM の値は同じになる．同じグループに含まれるデータどうしの類似度は考慮されていないため，元のデータと比較した情報損失を表していない．

3.3.3 情報エントロピーを用いた情報損失との比較

データセットのもつ情報の量を情報エントロピーとした場合の情報損失と ILD を比較する．情報エントロピーを用いた情報損失は，ILD と同様に，データの置き換えの前後での情報エントロピーの差の割合により計算するものとする．

データ間距離として離散距離を設定した場合，情報容量は値が異なる 2 つのデータの組み合わせを数え上げているものである．したがって，この場合の ILD は 2 つのデータが区別できなくなる割合を表しており，匿名化の効果を表す一つの指標として用いることができる．

例として，$N = nk$ 個の異なるデータがあり，それらを k 個ずつ n 個のグループに分割して，グループ内のデータをグループの代表値に置き換える匿名化を考える．グループの数 n を固定して k が十分大きい場合を考えると，最初はどの 2 つも区別可能であったデータが，匿名化後は同じグループに属する場合に区別不可能になるのであるから，その割合は約 $1/n$ である．離散距離に対する ILD は，この割合を測っているものである．実際，匿名化前の情報容量は，

$$I(X) = N(N-1)$$

で，分割後の情報容量は

$$I(\widehat{X}) = k^2 \cdot n(n-1) = N(N-k)$$

であるから，

$$ILD = \frac{k-1}{N-1} = \frac{k-1}{nk-1}$$

となり，k が大きいとき ILD$\sim 1/n$ である．

上の例において，データの生起確率はすべて $1/N$ であるものとして考えると，匿名化前の情報エントロピーは，

$$H(X) = \log N = \log n + \log k$$

で，匿名化後の情報エントロピーは

$$H(\widehat{X}) = \log n$$

であるから，情報エントロピーに関する情報損失は

$$\frac{H(X) - H(\widehat{X})}{H(X)} = \frac{\log k}{\log N} = \frac{\log k}{\log n + \log k}$$

となる．グループの数 n を固定して考えると，情報エントロピーに関する情報損失は，k が大きくなるに従いゆっくり 1 に近づく．これは，データセットに含まれる個々のデータの識別情報が徐々に失われていくことを表している．

このように，離散距離に関する ILD は 2 つのデータがどれぐらい区別できなくなるかを測るのに対し，情報エントロピーに関する情報損失は個別のデータがどれぐらい識別できなくなるかを測っており，情報損失指標として性質が大きく異なっている．匿名化のような応用分野では，個々のデータが識別，特定できなくなる度合いは重要な評価項目である．k-匿名化は，データセットに含まれるどのデータも k 人と区別がつかないようにデータセットを加工する手法であり，その評価においては，識別可能性に加えて区別可能性も考慮する必要がある．したがって，ILD は k-匿名化に対して有効な情報損失指標である．

第4章　ILDの適用と応用

ILDは，数値属性と非数値属性が両方含まれるデータセットへも容易に適用可能である．さらに，ILDはグループに含まれるデータどうしの類似度を反映した情報損失を表している．また，ILDは理論的に必ず情報損失を極小にする匿名化アルゴリズムの開発へ応用することができる．

本章の構成は次の通りである．4.1節では，数値属性と非数値属性が両方含まれるデータセットを匿名化し，その結果についてILDを計算し，ILDの適用について考察する．4.2節では，情報損失指標に基づき，アルゴリズム中の判定条件を決定し，情報損失を極小にする匿名化アルゴリズムの構成の例を説明する．

4.1　実データへのILDの適用

4.1.1　実験データセット

ILDを適用するデータセットとして，アメリカの国勢調査の結果を抽出したUCI adult dataset[57]を使用する．このデータセットのデータ総数は32,561である．また，含まれる属性数は14で，そのうち数値属性（整数値）が5個，非数属性（カテゴリデータ）が9個である．含まれる属性を表4.1に示す．

本実験では，数値データの属性であるcapital gainと，カテゴリデータの属性であるmarital statusを取り出し，2属性のデータ対のデータセットに対してILDを計算する．capital gainは0から99,999の整数値をとり，marital statusは7つのカテゴリデータ（Divorced，Married-AF-spouse，Married-civ-spouse，Married-spouse-absent，Never-married，Separated，Widowed）のいずれかの値をとる．

4.1.2　ミクロアグリゲーションの方法

ミクロアグリゲーションは，1. 分割，2. 集約（同一の値へデータを置き換える）の2つのステップにより実現される．

表 4.1: UCI adult dataset に含まれる属性

属性	種別
age	数値
workclass	8 種
fnlwgt	数値
education	16 種
education-num	数値
marital-status	7 種
occupation	14 種
relationship	6 種
race	5 種
sex	2 種
capital-gain	数値
capital-loss	数値
hours-per-week	数値
native-country	41 種

まず，データセットに対して異なる分割を行い，3 種類の分割されたデータセット D_A，D_B，D_C を生成する．分割の方法は，次の通りである．

- D_A は，数値データに関してより近い値どうしでグループをつくる分割である．データセットを数値データについて昇順になるようにソートし，小さい値から順に k 個ずつのグループを作り，残りのデータが $2k-1$ 個以下になったらそれらを 1 つのグループとする．このようにして分割されたデータセットを D_A とする．

- D_B は，カテゴリデータに関してより近い値どうしでグループをつくる分割である．カテゴリデータの文字列について昇順になるようにソートし，同様に k 個ずつのグループをつくる．このようにして分割されたデータセットを D_B とする．

- D_C は，どちらの属性についてもより近い値どうしでグループを作る分割である．まず，カテゴリデータの文字列について昇順になるようにソートし，次に同じカテゴリデータをもつデータを数値について昇順になるようにソートす

る．そして同様に k 個ずつのグループをつくる．このようにして分割されたデータセットを D_C とする．

次に，分割されたデータセット D_A，D_B，D_C について，それぞれグループ内のデータを同一の値に置き換える．同じグループに含まれるデータを置き換える代表値は，数値データはグループの平均値，カテゴリデータはグループ内で現れる頻度が最も高いデータとする．

4.1.3 ILD の適用結果

ILD の計算における距離の設定は，数値データのデータ間距離はユークリッド距離，カテゴリデータのデータ間距離は離散距離，データ組間の距離はユークリッド距離と離散距離の直積距離とした．また，直積距離の重みは，それぞれの属性の情報容量の逆数とした．D_A，D_B，D_C について，数値属性のみの ILD，カテゴリ属性のみの ILD，データセット全体の ILD を計算し，その結果を図 4.1~4.3 に示す．

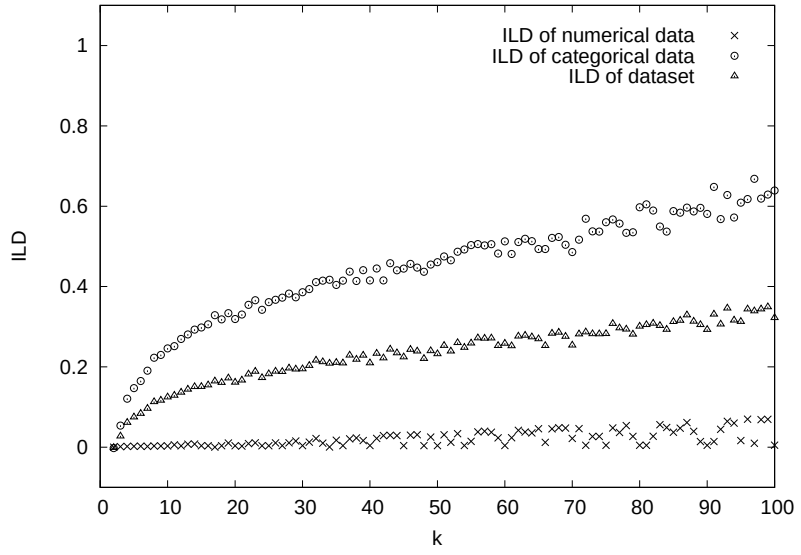

図 4.1: D_A の ILD

また，数値データのデータ間距離はユークリッド距離，カテゴリデータのデータ間距離は標準化されたレーベンシュタイン距離，データ組間の距離はユークリッド距

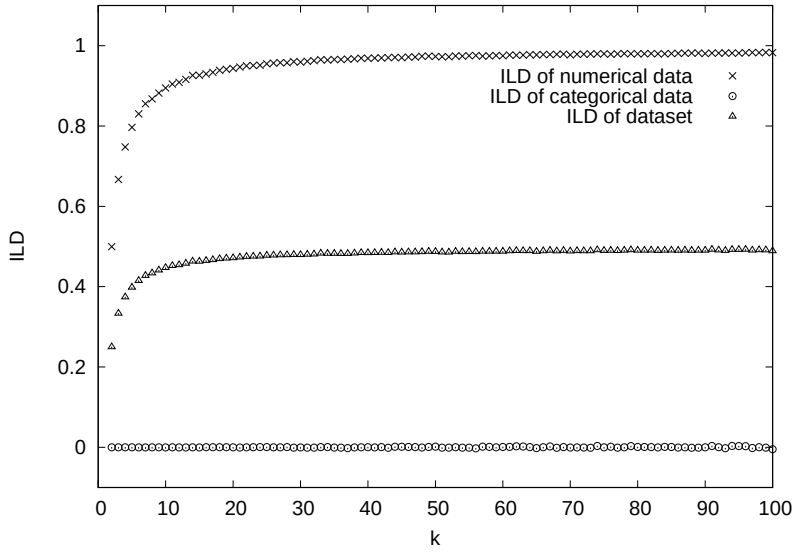

図 4.2: D_B の ILD

離と標準化されたレーベンシュタイン距離の直積距離とした場合についても計算した．また，直積距離の重みは，それぞれの属性の情報容量の逆数とした．D_A，D_B，D_C について，数値属性のみの ILD，カテゴリ属性のみの ILD，データセット全体の ILD を計算し，その結果を図 4.4~4.6 に示す．

4.1.4 ILD の適用に関する考察

D_A は，数値データについて近い値どうしを同じグループにしたため数値属性の ILD は小さいが，カテゴリデータについては分割の際に値を考慮されていないため，同じグループに含まれるカテゴリデータの値はばらばらとなり，k が大きくなるにつれてカテゴリ属性の ILD は大きくなる．このため，データセット全体の ILD は，k が大きくなるにつれて大きくなる．D_B についても同様の結果となっている．両方の属性の値を考慮して分割した D_C は，いずれの属性の ILD も大きく増加しないため，データセット全体の ILD は D_A，D_B と比較して小さくなっている．

また，文字列データについては，類似度を表す距離を設定することで，複数のデータを含むデータセットに対してもデータ間距離を自動的に与えることが可能となる．

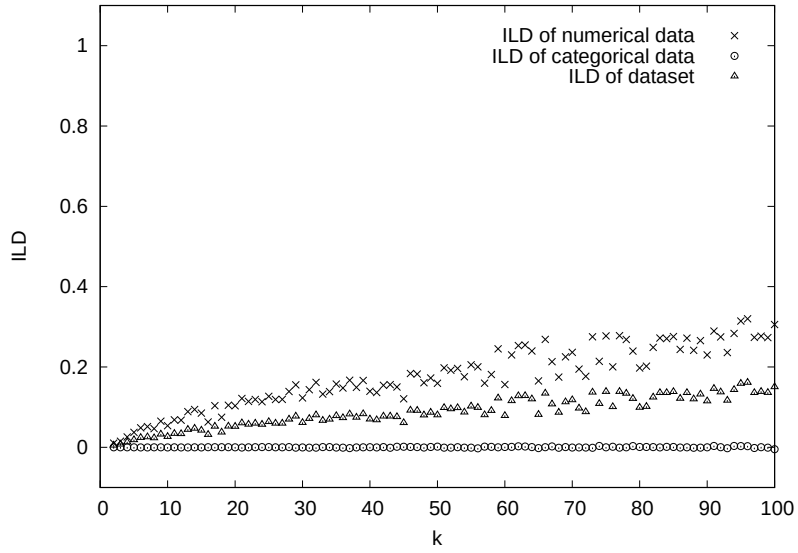

図 4.3: D_C の ILD

したがって，レーベンシュタイン距離などを用いれば，前項のように個々のデータ間の距離を個別に与えることなく，ILD を容易に適用することが可能である．

ILD を適用することにより，数値属性と非数値属性の両方を含むデータセットにも，グループ内のデータの類似度を反映した情報損失を求めることが可能である．さらに，匿名化されたデータセットは，元のデータと比較してどの程度情報を損失したかを表すことができる．これにより，より似た値どうしが同じグループに含まれているかというような，情報損失に関するミクロアグリゲーションの評価が可能となる．

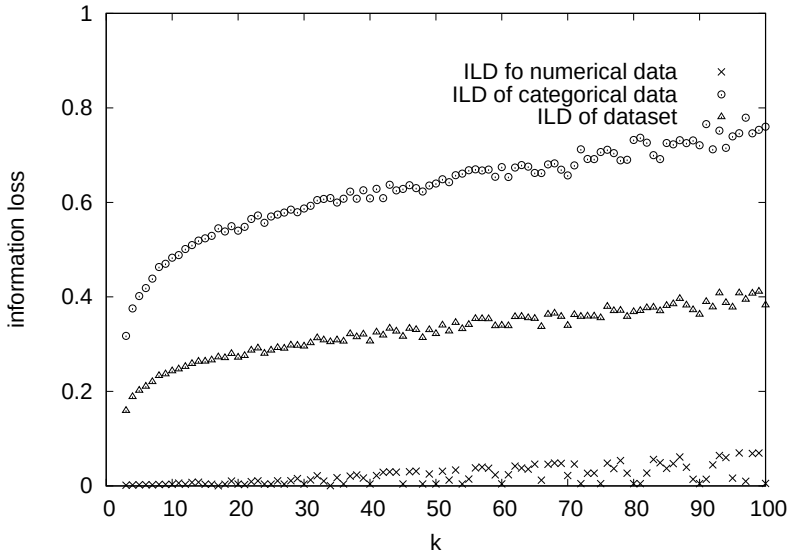

図 4.4: D_A のレーベンシュタイン距離を用いた ILD

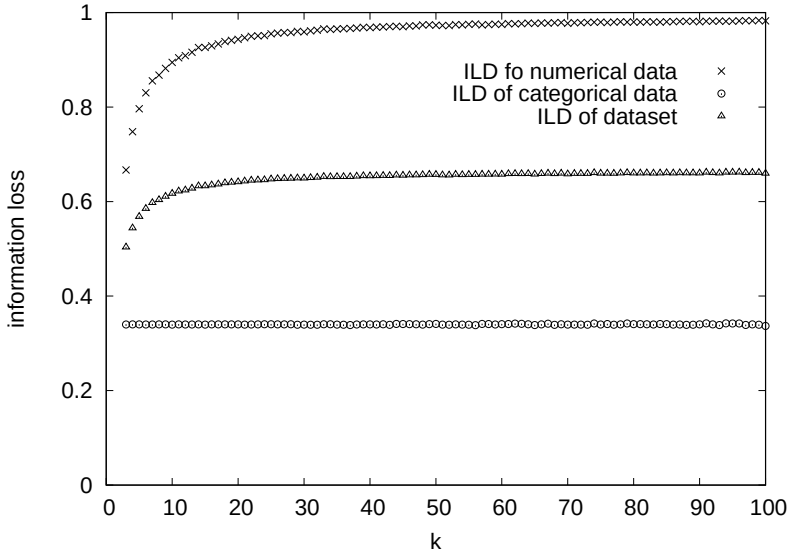

図 4.5: D_B のレーベンシュタイン距離を用いた ILD

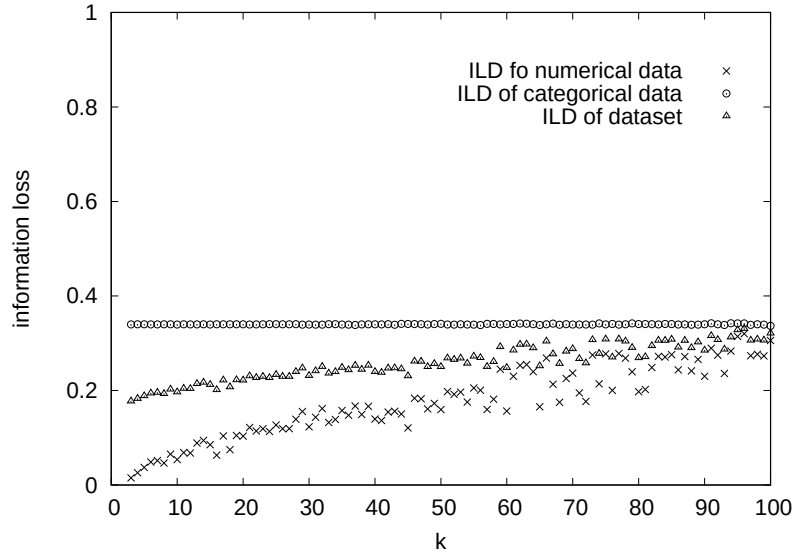

図 4.6: D_C のレーベンシュタイン距離を用いた ILD

4.2 匿名化アルゴリズムへの応用

4.2.1 情報損失を極小にするアルゴリズムの開発

匿名性を上げるとプライバシはより保護されるが，情報の量はより減少してしまうため，匿名性と有用性のバランスが問題となる．k-匿名性を情報に関しては，k-匿名性を維持しながら匿名化による情報の損失を抑えるアルゴリズムについて考える必要がある．ミクロアグリゲーションの手順の分割のための k 分割を行うアルゴリズムは存在しているが，情報損失の観点から必ずしも最良のものとはなっていない．

アルゴリズムの手順の考案に，情報損失指標の計算式を用いれば，情報損失を極小にすることが可能である．値の置き換えや削除といったデータの操作によって発生する情報損失の値を式で表し，その値が減少する場合にその操作を行うようにアルゴリズムを構成する．そうすれば，その操作を行うという条件においては，具体的な実データを用いて検証することなく，理論的に情報損失が極小となることを保証できる．さらに，有限回の操作で必ず終了することも保証できる．

本節では，情報損失指標を用いて情報損失が極小となるように構成したアルゴリズムの例として，MIL（Minimizing Information Loss）[58] の概要を説明する．MIL は，既存の k 分割アルゴリズムを用いて生成された匿名化情報を，k-匿名性を保持ししたまま分割を修正し，情報の損失を極小にするものである．MIL は，k-匿名性を満たす分割がされているデータセットについて，あるグループのデータを他のグループへ移動させることにより，情報損失を極小化させる (図 4.7)．

4.2.2 情報損失指標を用いた判定条件

アルゴリズム MIL の開発において，情報損失指標を用いたデータ移動の判定条件について説明する．

まず，MIL の実行において前提とする条件を示す．MIL にて取り扱うデータは，属性数が 1 の数値データとする．アルゴリズムの事前処理として，データを昇順にソートしているものとする．匿名化はミクロアグリゲーションとし，匿名化情報としてはクラスタの平均値を代表値とする．

D_i は $n_i(\geq k)$ 個のデータを含んでいるとする．ミクロアグリゲーションアルゴリズムを適用して得られる分割 $D_1, \ldots, D_g$ について，

$$D_i\text{の最大データ} \leq D_{i+1}\text{の最小データ} \tag{4.1}$$

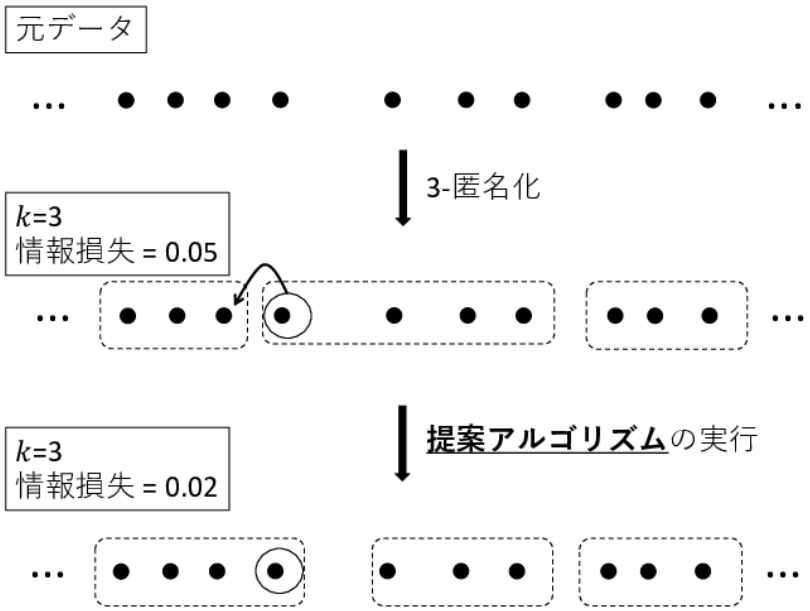

図 4.7: アルゴリズム MIL の概要

という条件がすべての i で成り立つ．また，情報損失の定義は，クラスタ内のデータと平均値との誤差二乗和の総和とする．

次に，あるクラスタに含まれる一つのデータを他のクラスタへ移動するとき情報損失が減少する条件式を示す．グループの代表値をグループの平均値とする場合，ILD は ILSSDM と同じ値になるため，ここでは ILSSDM の式を用いて説明する．あるクラスタのデータを $x_1, \ldots, x_n$，平均を $\overline{x}$ とし，誤差 2 乗和を

$$SSE_x = \sum_{j=1}^{n} (x_j - \overline{x})^2$$

とする．このとき，任意の a に対し，

$$\begin{aligned}\sum_{j=1}^{n}(x_j-a)^2 &= \sum_{j=1}^{n}x_j^2 - 2an\overline{x} + na^2 \\ &= \sum_{j=1}^{n}(x_j-\overline{x})^2 + n(\overline{x}-a)^2 \\ &= SSE_x + n(\overline{x}-a)^2 \end{aligned} \tag{4.2}$$

が成り立つ．このクラスタに新しいデータ x を追加したとき，$x_1,\ldots,x_n,x$ の平均 $\overline{x}'$ は，

$$\overline{x}' = \frac{1}{n+1}(\sum_{j=1}^{n}x_j + x) \tag{4.3}$$

$$= \frac{1}{n+1}(n\overline{x}+x) \tag{4.4}$$

となり，誤差 2 乗和 SSE'_x は (4.2) を用いて，

$$\begin{aligned}SSE'_x &= \sum_{j=1}^{n}(x_j-\overline{x}')^2 + (x-\overline{x}')^2 \\ &= SSE_x + n(\overline{x}-\overline{x}') + (x-\overline{x}')^2 \end{aligned} \tag{4.5}$$

となる．したがって，

$$SSE'_x - SSE_x = n(\overline{x}-\overline{x}')^2 + (x-\overline{x}')^2 \tag{4.6}$$

となる．ここで，式 (4.4) より，

$$\begin{aligned}SSE'_x - SSE_x &= n(\overline{x} - \frac{1}{n+1}(n\overline{x}+x))^2 + (x - \frac{1}{n+1}(n\overline{x}+x))^2 \\ &= \frac{n}{(n+1)^2}(\overline{x}-x)^2 + \frac{n^2}{(n+1)^2}(x-\overline{x})^2 \\ &= \frac{n}{n+1}(x-\overline{x})^2 \end{aligned} \tag{4.7}$$

となる．一方，(4.4) より $\overline{x}'$ を $\overline{x}$ で解くと，

$$\begin{aligned} n\overline{x} &= (n+1)\overline{x}' - x \\ \overline{x} &= \frac{1}{n}((n+1)\overline{x}' - x) \end{aligned}$$

となり，(4.6) に代入すると，

$$\begin{aligned} SSE'_x - SSE_x &= n((\frac{1}{n}((n+1)\overline{x}' - x) - \overline{x}')^2 + (x - \overline{x}')^2 \\ &= n(\frac{\overline{x}'}{n} - \frac{x}{n})^2 + (x - \overline{x}')^2 \\ &= \frac{n+1}{n}(\overline{x}' - x)^2 \end{aligned} \tag{4.8}$$

となる．式 (4.7)，(4.8) より，式 (4.6) は，

$$\begin{aligned} SSE'_x - SSE_x &= \frac{n}{n+1}(\overline{x} - x)^2 \\ &= \frac{n+1}{n}(\overline{x}' - x)^2 \end{aligned} \tag{4.9}$$

となる．

いま，2 つのクラスタ $x_1, \ldots, x_n, x$ と $y_1, \ldots, y_m$ があり，第 1 クラスタのデータ x を第 2 クラスタに移す場合について考える．第 1 クラスタの平均を $\overline{x}'$，$\overline{y}$，誤差 2 乗和の合計は，$SSE'_x + SSE_y$ から，

$$SSE_x + SSE'_y = (SSE'_x - \frac{n+1}{n}(x - \overline{x}')^2) + (SSE_y - \frac{m}{m+1}(x - \overline{y})^2)$$

に変化する．すなわち，$SSE'_x + SSE_y$ と $SSE_x + SSE'_y$ の差分について，

$$-\frac{n+1}{n}(x - \overline{x'})^2 + \frac{m}{m+1}(x - \overline{y})^2 < 0 \tag{4.10}$$

であれば，全体の SSE が減少する．したがって，式 (4.10) が成り立つ場合にデータの移動を行うという操作を繰り返すことにより，情報損失が極小である匿名化情報を生成することが可能となる．

4.2.3 アルゴリズム MIL

MIL にて行う操作と全体の流れについて説明する．

k-匿名性をみたす分割 $D_1, \ldots, D_g$ において条件 (4.1) が成り立っているとする．このとき，D_i，D_{i+1} の間で 1 つのデータを移動することを考える．

(a) D_i, D_{i+1} のデータを

$$D_i = \{x_1, \ldots, x_n, x\}$$
$$D_{i+1} = \{y_1, \ldots, y_m\}$$

と表すこととし，D_i の最大データが x，平均が $\overline{x}'$，D_{i+1} の平均が $\overline{y}$ とする．このとき，$n \geq k$ かつ

$$-\frac{n+1}{n}(x - \overline{x}')^2 + \frac{m}{m+1}(x - \overline{y})^2 < 0 \tag{4.11}$$

ならば，データを移動したほうが情報損失が小さくなるため，x を D_i から D_{i+1} に移動する．この操作を条件がなりたつ限り繰り返す．

(b) 条件 (4.11) がなりたたない場合，

$$D_i = \{x_1, \ldots, x_n\}$$
$$D_{i+1} = \{x, y_1, \ldots, y_m\}$$

とし，D_i の平均が $\overline{x}$，D_{i+1} の最小データが x，平均が $\overline{y}'$ とする．このとき，$m \geq k$ かつ

$$-\frac{n}{n+1}(x - \overline{x})^2 + \frac{m+1}{m}(x - \overline{y}')^2 > 0 \tag{4.12}$$

ならば x を D_{i+1} から D_i に移動する．この操作を条件がなりたつ限り繰り返す．

(a), (b) の処理をすべての隣り合うクラスタ D_i, D_{i+1} $(i = 1, \ldots, g-1)$ 間で行う．データの移動が一度でも発生した場合，再び先頭のクラスタから同様の処理を行う．これらの処理を繰り返し，どの隣り合うクラスタについてもデータの移動がなかった場合，アルゴリズムを終了する．上記のアルゴリズムを Algorithm 1 に示す．ここで，$|D_i|$ はクラスタ D_i のデータ数を表す．

実行時間について，データセットとは別に書くクラスタの要素数と平均を保持するものとすれば，条件 (4.11)，(4.12) の判定とデータ移動があった際の要素数や平均の更新は，データ数によらず一定時間に効率よく行える．

このアルゴリズムによって得られた分割は，クラスタの 1 つの要素を別のクラスタへ移すことによって，これ以上情報損失を減少させることができないという意味で極小ではあるが，2 個以上の要素の同時移動やクラスタ数の変更によりさらに情報損失が減少する可能性があるので，必ずしも情報損失が最小になるとは限らないことに注意する．

Algorithm 1 MIL (minimizing information loss)

$flag \leftarrow 1$
while $flag = 1$ **do**
  $flag \leftarrow 0$
  for $i = 1$ to $g - 1$ **do**
    while $|D_i| > k$ **do**
      $n = |D_i| - 1$
      $m = |D_{i+1}|$
      $X = -\frac{n+1}{n}(x - \overline{x}')^2 + \frac{m}{m+1}(x - \overline{y})^2$
      if $X < 0$ **then**
        $D_i = \{x_1, \cdots, x_n\}$
        $D_{i+1} = \{x, y_1, \cdots, y_m\}$
        $flag \leftarrow 1$
      else
        break
      end if
    end while
    while $|D_{i+1}| > k$ **do**
      $n = |D_i|$
      $m = |D_{i+1}| - 1$
      $X = -\frac{n}{n+1}(x - \overline{x})^2 + \frac{m+1}{m}(x - \overline{y}')^2$
      if $X > 0$ **then**
        $D_i = \{x_1, \cdots, x_n, y\}$
        $D_{i+1} = \{y_1, \cdots, y_m\}$
        $flag \leftarrow 1$
      else
        break
      end if
    end while
  end for
end while

第5章　結論

本研究では，匿名化によりデータセットのデータの値が置き換えられたときに失われる情報を測ることを目的とし，データ間距離に基づいて計算する情報損失指標 ILD を考案した．ILD の定義のため，データセットのもつ情報の量を表す情報容量を定義した．これは，データ間の距離に基づき計算できるものである．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．また，非数値データについては，距離を設定することによりデータどうしの類似度を表すことが可能となった．さらに，これまでは数値データと非数値データを両方含むデータセットに対しては，情報損失を測るのは難しかったが，距離の設定として直積距離を考えることで簡単に適用が可能となった．したがって，ILD は，多くのデータセットへ適用することが可能であるため，匿名化の評価に広く活用することができるといえる．

実際のデータセットへの適用の例として，アメリカの国勢調査のデータセットに対して，異なる複数のミクロアグリゲーションを実行し，情報損失を計算した．その結果，より似ているデータが同じグループに含まれる分割が最も ILD の値が小さくなった．したがって，ILD を用いることにより，複数の異なる種別の属性を含むデータセットに対しても，似ているデータどうしが同じグループに含まれているかどうかが測れるようになった．

さらに，ILD は匿名化アルゴリズムの開発に応用することが可能である．分割などの匿名化の操作により発生する情報損失は，ILD の式を用いて表すことができる．したがって，匿名化アルゴリズムは，ある操作を行うことで情報損失が減少するという条件を ILD を用いた判定式として表すことにより，その操作を行う限りにおいては情報損失が極小となるアルゴリズムを開発することができる．

謝辞

学位取得に向けて，あたたかいお言葉やお心遣いをいただきました，奈良女子大学の加古富志雄教授，城和貴教授，山下靖教授に心から感謝いたします．

研究指導や論文執筆等，数多くのご助言や丁寧で熱心なご指導，あたたかい励ましをいただきました，大阪大学大学院情報科学研究科の和田昌昭教授に心から感謝いたします．

最後に，いつも応援してくれた家族に心から感謝いたします．

本研究では，提案アルゴリズム MIL の適用において，Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. を利用しました．

参考文献

[1] 個人情報の保護に関する法律. http://law.e-gov.go.jp/htmldata/H15/H15HO057.html.

[2] プライバシー保護と個人データの国際流通についてのガイドラインに関する oecd 理事会勧告. http://www.mofa.go.jp/mofaj/gaiko/oecd/privacy.html.

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

[4] 高度情報通信ネットワーク社会推進戦略本部（ＩＴ総合戦略本部），パーソナルデータに関する検討会，資料 2-1 技術検討ワーキンググループ報告書. http://www.kantei.go.jp/jp/singi/it2/pd/dai5/siryou2-1.pdf.

[5] 個人情報の保護に関する法律 (平成 15 年法律第 57 号. https://www.ppc.go.jp/files/pdf/personal_law.pdf.

[6] 個人情報保護法改正で重要性増す匿名化技術とその進化. http://www.toshiba.co.jp/iot/power/entry/2015/2015_004.htm.

[7] 高橋克己. 匿高度情報通信ネットワーク社会推進戦略本部（ＩＴ総合戦略本部），パーソナルデータに関する検討会，資料 2-3 技名化技術の現状について. http://www.kantei.go.jp/jp/singi/it2/pd/wg/dai1/siryou2_3.pdf.

[8] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557–570, 2002.

[9] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, Vol. 1, No. 1, March 2007.

[10] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pp. 106–115. IEEE, 2007.

[11] Traian Marius Truta and Bindu Vinay. Privacy protection: p-sensitive k-anonymity property. In *ICDE workshops*, p. 94. Citeseer, 2006.

[12] 五十嵐大, 千田浩司, 高橋克巳. k-匿名性の確率的指標への拡張とその応用例. *CSS*, 2009.

[13] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 49–60. ACM, 2005.

[14] 原田邦彦, 佐藤嘉則. 一般化階層木の自動生成と情報エントロピーによる歪度評価を伴う k-匿名化手法. 情報処理学会研究報告. CSEC,[コンピュータセキュリティ], Vol. 2010, No. 47, pp. 1–7, 2010.

[15] Latanya Sweeney. Guaranteeing anonymity when sharing medical data, the datafly system. In *Proceedings of the AMIA Annual Fall Symposium*, p. 51. American Medical Informatics Association, 1997.

[16] Anco Hundepool and LCRJ Willenborg. μ-and τ-argus: Software for statistical disclosure control. In *Third International Seminar on Statistical Confidentiality*, 1996.

[17] 村本俊祐, 上土井陽子, 若林真一. k-匿名性を利用したデータ一般化によるプライバシー保護. In *DEWS*, 2007.

[18] 村本俊祐, 上土井陽子, 若林真一. データを極小歪曲し k-匿名性を保持したデータに変換するプライバシー保護アルゴリズム. *DBSJ Letters*, Vol. 6, No. 1, 2007.

[19] 村本俊祐, 上土井陽子, 若林真一. 背景知識を用いた推測を困難にしデータ歪曲度を極小化するプライバシー保護手法. *DEWS2008 C1-4*.

[20] 渡邉奈津美, 土井洋, 趙晋輝. k-匿名化手法の効率向上に関する一提案. 情報処理学会第 75 回全国大会, Vol. 2, p. 2, 2013.

[21] Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology (JOPT)*, 2005.

[22] Hyoungmin Park and Kyuseok Shim. Approximate algorithms for k-anonymity. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 67–78. ACM, 2007.

[23] Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 223–228. ACM, 2004.

[24] Gagan Aggarwal, Tomás Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Anonymizing tables. In *Database Theory-ICDT 2005*, pp. 246–258. Springer, 2005.

[25] 木村映善. k-匿名性を利用した医療保健情報の利用可能性についての考察: 国内外の医療情報利用に関する事例から (セキュリティシステム, インターネットと情報倫理教育, 一般). 電子情報通信学会技術研究報告. SITE, 技術と社会・倫理, Vol. 111, No. 484, pp. 223–228, 2012.

[26] Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, et al. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, Vol. 16, No. 5, pp. 670–682, 2009.

[27] 荒牧英治, 増川佐知子, 宮部真衣, 森田瑞樹. テキストの k-匿名化. 情報処理学会研究報告. データベース・システム研究会報告, Vol. 2012, No. 9, pp. 1–8, 2012.

[28] 高橋翼, 宮川伸也, 伊東直子. 移動軌跡ストリームに対するリアルタイム k 匿名化手法の提案, 2011.

[29] 清雄一, 大須賀昭彦. 誤差を考慮した位置匿名化手法の提案. 電子情報通信学会論文誌 D, Vol. 97, No. 5, pp. 964–974, 2014.

[30] 川本淳平, 福地一斗, 照屋唯紀, 佐久間淳. プライバシを考慮した移動系列情報解析のための安全性の提案. *SCIS 2013*, 2013.

[31] 村上啓介, 宇野毅明. 大規模データに対する情報損失の少ない k-匿名化手法. 情報処理学会研究報告. 情報学基礎研究会報告, Vol. 2013, No. 34, pp. 1–6, 2013.

[32] 村上啓介, 宇野毅明. マッチングアルゴリズムを用いた大規模データ k-匿名化の解法. 情報処理学会研究報告. AL, アルゴリズム研究会報告, Vol. 2013, No. 8, pp. 1–8, 2013.

[33] 青木俊介, 岩井将行, 瀬崎薫. 参加型環境センシングを用いた統計情報構築のためのプライバシー保護手法. 電子情報通信学会論文誌 B, Vol. 97, No. 1, pp. 41–50, 2014.

[34] 山岡裕司, 伊藤孝一, 牛田芽生恵, 津田宏. プライバシー保護データ開示における 2-多様性を満たす機微データ曖昧化法. 電子情報通信学会論文誌 A, Vol. 97, No. 3, pp. 197–208, 2014.

[35] 菊池亮, 山中章裕, 五十嵐大. プライバシー保護されたデータに対する t 検定手法 (セキュリティ, ライフログ活用技術, オフィスインフォメーションシステム, ライフインテリジェンス, 一般). 電子情報通信学会技術研究報告. LOIS, ライフインテリジェンスとオフィス情報システム, Vol. 111, No. 470, pp. 171–176, 2012.

[36] 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*, 2006.

[37] Josep Domingo-Ferrer and Josep Maria Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and data Engineering*, Vol. 14, No. 1, pp. 189–201, 2002.

[38] Vicenç Torra. Microaggregation for categorical variables: a median based approach. In *International Workshop on Privacy in Statistical Databases*, pp. 162–174. Springer, 2004.

[39] Ji-Won Byun, Ashish Kamra, Elisa Bertino, and Ninghui Li. Efficient k-anonymization using clustering techniques. In *Advances in Databases: Concepts, Systems and Applications*, pp. 188–200. Springer, 2007.

[40] Michael Laszlo and Sumitra Mukherjee. Minimum spanning tree partitioning algorithm for microaggregation. *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 17, No. 7, pp. 902–911, 2005.

[41] Jun-Lin Lin and Meng-Cheng Wei. An efficient clustering method for k-anonymization. In *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*, pp. 46–50. ACM, 2008.

[42] Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Samir Khuller, Rina Panigrahy, Dilys Thomas, and An Zhu. Achieving anonymity via clustering. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 153–162. ACM, 2006.

[43] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pp. 25–25. IEEE, 2006.

[44] Tochukwu Iwuchukwu and Jeffrey F Naughton. K-anonymization as spatial indexing: Toward scalable and incremental anonymization. In *Proceedings of the 33rd international conference on Very large data bases*, pp. 746–757. VLDB Endowment, 2007.

[45] Jerome H Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software (TOMS)*, Vol. 3, No. 3, pp. 209–226, 1977.

[46] Antonin Guttman. *R-trees: a dynamic index structure for spatial searching*, Vol. 14. ACM, 1984.

[47] Anna Oganian and Josep Domingo-Ferrer. 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.

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

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

[50] 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.

[51] RC Jancey. Multidimensional group analysis. *Australian Journal of Botany*, Vol. 14, No. 1, pp. 127–130, 1966.

[52] 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.

[53] 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.

[54] Vijay S Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 279–288. ACM, 2002.

[55] Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *21st International Conference on Data Engineering (ICDE'05)*, pp. 217–228. IEEE, 2005.

[56] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. Utility-based anonymization using local recoding. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 785–790. ACM, 2006.

[57] M. Lichman. UCI machine learning repository, 2013. `http://archive.ics.uci.edu/ml`.

[58] 秋山寛子, 和田昌昭, 中山雅哉, 加藤朗, 砂原秀樹. k-匿名化アルゴリズムにおける情報損失の極小化. 情報処理学会論文誌, Vol. 57, No. 12, pp. 2675–2681, dec 2016.

[59] 脇本和昌. 乱数の知識. 森北出版, 1970.

[60] 平岡和幸, 堀玄. プログラミングのための確率統計. オーム社, 2009.

研究業績

学位に関わる論文

学術論文

1. 秋山 寛子，和田 昌昭：データ間の距離に基づく情報損失指標，情報処理学会論文誌「数理モデル化と応用」，Vol.11，No.3，pp.100-105 （2018-12-20）

その他の業績

学術論文

1. 秋山 寛子，和田 昌昭，中山 雅哉，加藤 朗，砂原 秀樹 : 匿名化アルゴリズムにおける情報損失の極小化，情報処理学会論文誌, Vol.57, No.12, pp.2675-2681 (2016-12-15)

国内学会【査読あり】

1. 秋山 寛子，山内 正人，落合 秀也，砂原 秀樹：センサネットワーク基盤技術 IEEE1888 を用いた実装と評価，マルチメディア，分散協調とモバイルシンポジウム，(2012.7)
2. 秋山 寛子，山内 正人，柴崎 亮介，砂原 秀樹：情報銀行システムにおける個人情報蓄積機構の機能設計と実装，マルチメディア，分散協調とモバイルシンポジウム 2013，pp.1953-1957，(2013.7)
3. 辻井 高浩，山内 正人，秋山 寛子，猪俣 敦夫，藤川 和利，砂原 秀樹 : 東大寺内におけるネットワーク整備について，マルチメディア，分散協調とモバイルシンポジウム 2013，pp.435-441，(2013.7)

4. 秋山寛子，中山雅哉，加藤朗，砂原秀樹：2 つの匿名化情報の組み合わせによる k-匿名度の定式化に関する考察，マルチメディア，分散，強調とモバイルシンポジウム 2014，pp.234-240(2014.7)

5. 芦田 和毅，秋山 寛子，市川 敬夫：MPU の作成を題材とした HDL 学習教材の開発，平成 27 年度工学・工業教育研究講演会，1G06，pp.142-143，(2015.9)

6. 芦田 和毅，秋山 寛子，村田 雅彦，藤澤 義範：電子天秤作成による電子回路の実装可能な技術者の育成教育，平成 28 年度工学教育研究講演会，1G07，pp.134-135，(2016.9)

国内学会【査読なし】

1. 秋山 寛子，和田 昌昭：ラスター画像から面積を求めるための閾値自動決定アルゴリズム，バイオイメージング学会,（2009.9）

2. 秋山 寛子，加藤 朗，砂原 秀樹：匿名化情報のマッチングにおける匿名度の定式化についての一考察，電子情報通信学会技術研究報告書，情報通信システムセキュリティ 113.502，pp167-172，(2014.3)

3. 秋山 寛子，加藤 朗，砂原 秀樹：再識別可能性を考慮した匿名度指標の一検討，マルチメディア情報ハイディング・エンリッチメント研究会，信学技報，vol.115，no.479，EMM2015-76，pp.1-5，(2016.3)

4. 秋山 寛子，和田 昌昭，中山 雅哉，加藤 朗，砂原 秀樹：データ間距離に基づいた匿名化アルゴリズムの一考察，電子情報通信学会総合大会，vol.2016，no.2，pp.192，(2016.3)

5. 秋山 寛子，和田 昌昭：情報損失指標の非数値データへの適用，情報処理学会数理モデル化と問題解決研究会,（2017.6）

6. 重岡 広大，秋山 寛子，和田 昌昭：k-匿名分割における元の交換による情報損失減少アルゴリズム，電子情報通信学会マルチメディア情報ハイディング研究会,（2017.11）

一般講演・その他

1. 秋山 寛子，松村 真紀子，加古 富志雄：素な非交代結び目の表の作成，奈良女子大学人間文化研究科年報，(2008.10)

2. Hiroko Akiyama, Hideya Ochiai, Akira Kato, Hideki Sunahara:Tutorial IEEE1888, APNG(Asia Pacific Networking Group)，(2012.8)

3. 秋山 寛子：パーソナル情報保護における匿名化手法の実験と考察，2013 年 WIDE 冬研究会，(2013.12)

4. 秋山 寛子：情報損失が極小となる匿名化アルゴリズムの開発，平成 27 年度善光寺バレー研究成果報告会，(2015.10)

5. 秋山 寛子，和田 昌昭：ミクロアグリゲーションにおける情報損失について，電子情報通信学会マルチメディア情報ハイディング研究会，(2018.3)

付 録A　MILの適用による情報損失の減少と実行時間

A.1　MIL適用による情報損失の減少

A.1.1　MIL適用による情報損失に関する評価の概要

本節では，既存のk匿名化アルゴリズムの分割に対して，MILを実行することにより，情報損失がどの程度減少するかを調査する．実験は以下の手順で行う．

1. 様々な確率分布に従う乱数データセットを生成する．
2. MDAV，VMDAVの分割結果に対してMILを実行し，MDAV+MIL，VMDAV+MILの分割を生成する．
3. 各分割の情報損失を比較する．

A.1.2　MIL適用に使用するデータセットの生成

表A.1に示す13種の異なる確率密度関数$f(x)$を設定し，それらの分布に従うデータセットを生成する．ただし，$N(\mu, \sigma)$は，平均μ，分散σ^2の正規分布を表す．データセットのうち，DS0は実データを表す最も一般的な分布である正規分布とした．また，MILによる情報損失の減少を確認するため，分布に偏りのあるデータセットDS1∼DS11を作成する．分布の偏りについて，複数箇所に均等に偏りがある場合と，複数箇所に異なる大きさの偏りがある場合について，確率密度関数を設定した．DS12は，無数に偏りがある場合を表す分布として一様分布とした．

まず，正規乱数のデータセットを生成する方法について述べる．一様分布データd個の平均値$\overline{X}$の確率分布は，中心極限定理により近似的に正規分布に従うので，サ

表 A.1: データセットの確率分布

データセット	確率密度関数 $f(x)$	データ総数
DS0	$N(0,1)$	100
DS1	$\frac{1}{2}N(5,1)+\frac{1}{2}N(10,1)$	200
DS2	$\frac{1}{2}N(5,1)+\frac{1}{2}N(8,1)$	200
DS3	$\frac{1}{2}N(5,1)+\frac{1}{2}N(10,2)$	200
DS4	$\frac{1}{2}N(10,3)+\frac{1}{2}N(20,2)$	200
DS5	$\frac{1}{3}N(0,1)+\frac{1}{3}N(5,2)+\frac{1}{3}(12,3)$	300
DS6	$\frac{1}{3}N(5,1.5)+\frac{1}{3}N(10,1)+\frac{1}{3}N(15,1.5)$	300
DS7	$\frac{1}{3}N(5,3)+\frac{1}{3}N(15,2)+\frac{1}{3}N(20,1)$	300
DS8	$\frac{1}{3}N(5,3)+\frac{1}{3}N(12,1.5)+\frac{1}{3}N(20,2)$	300
DS9	$\frac{1}{3}N(5,2)+\frac{1}{3}N(10,1.5)+\frac{1}{3}N(18,3)$	300
DS10	$\frac{1}{3}N(0,1)+\frac{1}{3}N(5,1)+\frac{1}{3}N(10,1)$	300
DS11	$\frac{1}{3}N(0,1)+\frac{1}{3}N(3,1)+\frac{1}{3}N(6,1)$	300
DS12	$1(0\leq x\leq 1),0(otherwise)$	100

ンプルデータは，

$$z(x)=\frac{\overline{X}-\frac{1}{2}}{\sqrt{\frac{1}{12d}}}$$

で算出できる ([59])．平均が μ，分散が σ^2 である確率分布 $f(x)$ の場合，サンプルデータは $z_i(x)=\sigma z(x)+\mu$ となる．このようにして算出したデータの集合を正規乱数のデータセットとする．本実験では，$d=6$ として正規乱数列を生成する．

次に，異なる正規乱数を重ね合わせた確率分布に従うサンプルデータセットの生成について述べる．w 個の確率分布を重ね合わせる場合，$f_1(x)$ に従うサンプルデータセットを $x_1,\ldots,x_{n_1}$，$f_2(x)$ に従うサンプルデータセットを $y_1,\ldots,y_{n_2}$ というように，各確率分布に対して生成したサンプルデータセットを合わせたデータセット $x_1,\ldots,x_{n_1},y_1,\ldots,y_{n_2},\ldots$ は，確率分布 $f(x)=\frac{1}{\sum_{i=1}^{w}n_i}(n_1f_1(x)+\cdots+n_wf_w(x))$ に従うサンプルデータセットである ([60])．本実験では，異なる平均と分散をもつ正規分布を重ねることにより，偏りのあるデータセットの生成を行う．

A.1.3 k-匿名化アルゴリズムと MIL の実行

前節で生成したデータセットを用いて，既存アルゴリズム MDAV，VMDAV と MIL を実行し，情報損失を比較する．既存アルゴリズムは，データセット及び k 値を与えて実行する．データセットを記述したファイルには，1 行に 1 つの実数のデータが書かれている．実行結果は，1 行に「データ クラスタ番号」を記述したものを出力する．クラスタ番号は，昇順に連番で付与する．なお，VMDAV の実行では，アルゴリズムのパラメータ γ を 1.0 とする．

MIL は，既存アルゴリズムにより分割されたデータセットと k 値を与えて実行する．データ総数を N とすると，k 値は $2 \sim \frac{N}{2}$ について実行する．分割されたデータセットとして，1 行に「データ クラスタ番号」を記述したものを読み込み，アルゴリズムを実行しクラスタ内のデータを調整し，クラスタ番号を更新する．データの移動について，隣り合う 2 つのクラスタに注目するが，4.11 の判定では 1 つ目のクラスタの始点と終点，2 つ目のクラスタの終点の 3 点を見つければ判定ができる．また，4.12 の判定では 1 つ目のクラスタの始点，2 つ目のクラスタの始点と終点の 3 点を見つければ良い．

情報損失は，数値に付与されたクラスタ番号を用いて算出する．また，クラスタの代表値はクラスタ内のデータの算術平均とする．

A.1.4 MIL による情報損失減少の検証

MDAV と VMDAV の実行結果の情報損失に対して，MDAV+MIL，VMDAV+MIL の実行結果の情報損失を比較する．実験結果の例として，標準正規分布である DS0，分布の偏りが同一である DS1，より複雑な分布の偏りをもつ DS9 について述べる．

まず，DS0 に関して，MDAV と MDAV+MIL の情報損失の結果を図 A.1，VMDAV と VMDAV+MIL の情報損失の結果を図 A.2 に示す．また，$k = 2 \sim 15$ の場合の情報損失を表 A.2 に示す． 次に，DS1 に関して，確率分布を図 A.3，MDAV と MDAV+MIL の情報損失の結果を図 A.4，VMDAV と VMDAV+MIL の情報損失の結果を図 A.5 に示す．また，$k = 2 \sim 15$ の場合の情報損失を表 A.3 に示す． 最後に，DS9 に関して，確率分布を図 A.6，MDAV と MDAV+MIL の情報損失の結果を図 A.7，VMDAV と VMDAV+MIL の情報損失の結果を図 A.8 に示す．また，$k = 2 \sim 15$ の場合の情報損失を表 A.4 に示す． その他のデータセットの情報損失の結果については，付録 A に示す．

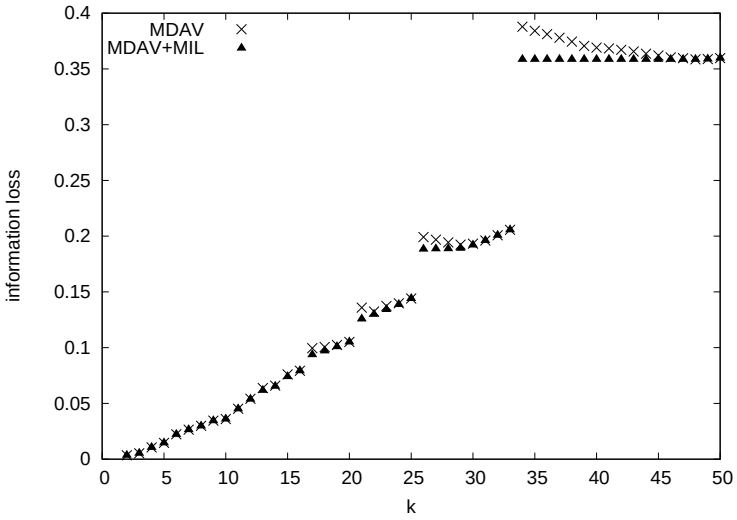

図 A.1: DS0 の情報損失の比較 (MDAV)

DS0∼DS12 に関して，MDAV については，全てのデータセットに対して $k = 2 \sim \frac{N}{2}$ のうち 66.5%が MIL により改善され，情報損失の減少率は平均 12.6%，最大 68.8%であった．また，VMDAV については 89.9%が改善され，情報損失の減少率は平均 8.9%，最大 51.7%であった．MDAV と VMDAV に対する情報損失の減少の結果を表 A.5，A.6 に示す．

A.1.5 MIL 適用による情報損失の減少についての考察

MIL による情報損失の減少は，データ総数に対する k の値という観点から整理することができる．k が小さい場合や，データ総数を k で割った余りが小さい場合は，MDAV 及び VMDAV の情報損失は，MIL による改善は少なかった．MIL はデータ数が k より大きいクラスタからのデータ移動を行うアルゴリズムであり，これらの場合はデータ移動の自由度が低いため情報損失の減少率が小さかったと考えられる．一方，データ総数を k で割った余りが大きい場合は，MIL による情報損失が大きく減少する場合が多く見られた．したがって，情報損失の減少はデータ総数に対する k の値に依存すると言える．

MDAV について，データ総数が k の倍数の場合は，MDAV によりデータ数 k のク

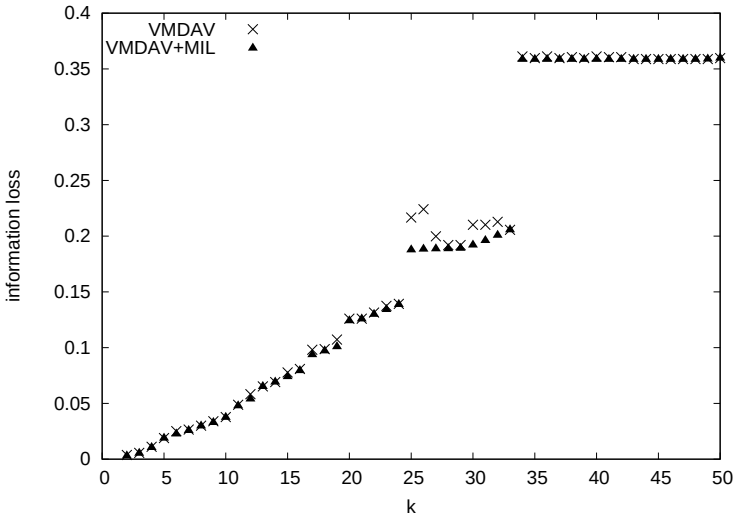

図 A.2: DS0 の情報損失の比較 (VMDAV)

ラスタに分割されるため，データ移動による情報損失の改善の余地はない．VMDAV はクラスタ数が必ずしも k で割った値（または k で割った値に 1 を加えたもの）とはならない．そのため，クラスタ内の要素数が大きくクラスタ数が小さくなるため，MDAV よりも情報損失が大きくなる場合が多くなり，MDAV と比較して MIL を行った結果の情報損失が大きい k が多く見られた．特に，$k = \frac{N}{2}$ のとき，クラスタ数が 1 になる場合があり，この場合，情報損失は 1 となり減少させることができない．

A.2 MIL の実行時間

A.2.1 MIL の実行時間に関する評価の概要

MIL の実行時間について，データ総数に対する実行時間の増分を調査する．実行時間は，アルゴリズムのループ内にてデータ移動を判定する回数を用いて評価する．実験の手順を以下に示す．

1. A.1.2 のようにして DS0∼DS11 の確率密度関数に従うデータ数 N が 100，1000，10000，100000 のデータセットを生成する．データセットは，乱数のシードを変えて 3 セットずつ用意する．

表 A.2: DS0 の情報損失の値

k	MDAV	MDAV+MIL	VMDAV	VMDAV+MIL
2	0.003446	0.003446	0.003765	0.002946
3	0.005258	0.005258	0.005632	0.005282
4	0.010552	0.010552	0.011109	0.010479
5	0.014560	0.014560	0.018915	0.018828
6	0.022300	0.022263	0.025131	0.022560
7	0.026470	0.026470	0.026578	0.025728
8	0.029978	0.029778	0.030077	0.029778
9	0.034523	0.034490	0.034029	0.032834
10	0.035952	0.035952	0.037727	0.037490
11	0.045240	0.045126	0.048749	0.047834
12	0.054165	0.053752	0.058318	0.053937
13	0.063774	0.061702	0.065301	0.065207
14	0.065962	0.065343	0.069149	0.069149
15	0.076020	0.074068	0.077760	0.074068

2. (1) にて生成したデータセットについて MDAV と VMDAV を実行し，クラスタを生成する．

3. (2) にて生成したクラスタに MIL を実行し，データ移動の判定回数を $k = 2 \sim 50$ についてカウントする．

A.2.2 データ総数に対する実行時間の調査結果

DS0 に関して，シードが 0，1，2 のデータセットについて判定回数をプロットした結果を図 A.9∼A.14 に示す．また，各データセットについて，$k = 2 \sim 50$ の範囲での判定回数の平均を表 A.7，最大を A.8 に示す．その他のデータセットについては付録 B に示す．

DS0∼DS12 に関して，判定回数の平均と最大値を表 A.9 にまとめる．

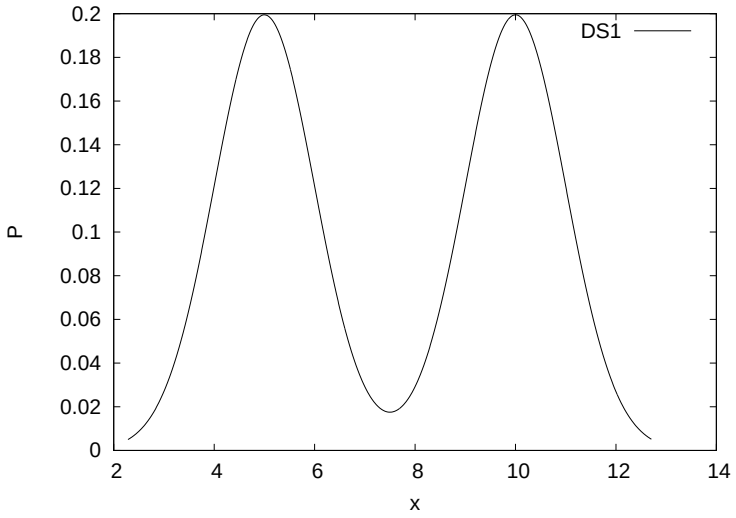

図 A.3: DS1 の確率分布

A.2.3 MIL の実行時間についての考察

正規分布の場合，MDAV による分割の結果，クラスタのデータ数は$k, \ldots, k, l, k, \ldots, k (k \geq l < 2k)$ のようになる．特に DS0 の正規分布の場合，MIL により，データ数 l のクラスタの $(l-k)$ 個以下のデータはその付近の $(l-k)$ 個以下のクラスタに移動する．判定が必要なのは，それらのクラスタのみである．したがって，判定回数は k のみに依存し，データ総数には依存しないため，データ総数が大きくなっても MIL は高速に実行可能である．DS1~DS12 についても同様の結果となり，データ総数が増えても実行時間の大きな増加は見られなかった．

VMDAV について，実験結果よりデータ数に比例して実行時間が増加しているため，実行時間は $O(N)$ であると考えられる．これは，VMDAV を実行するとクラスタ数が $\frac{N}{k}$ に対して最大約 8 割となっており，判定回数が増加したためであると思われる．

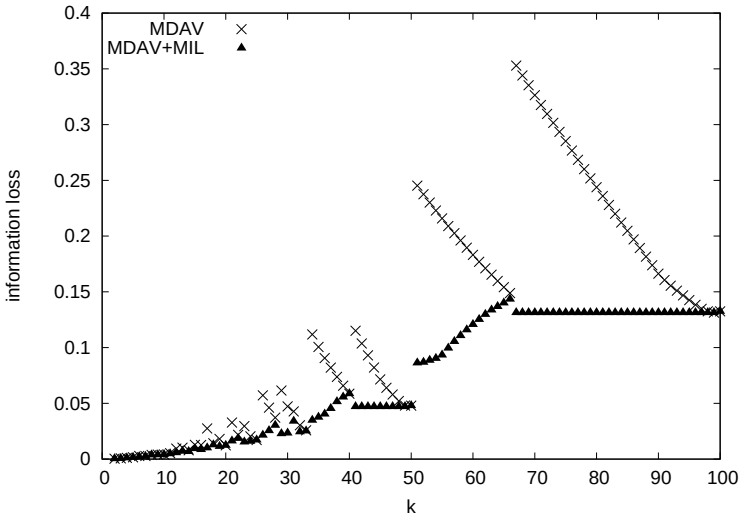

図 A.4: DS1 の情報損失の比較 (MDAV)

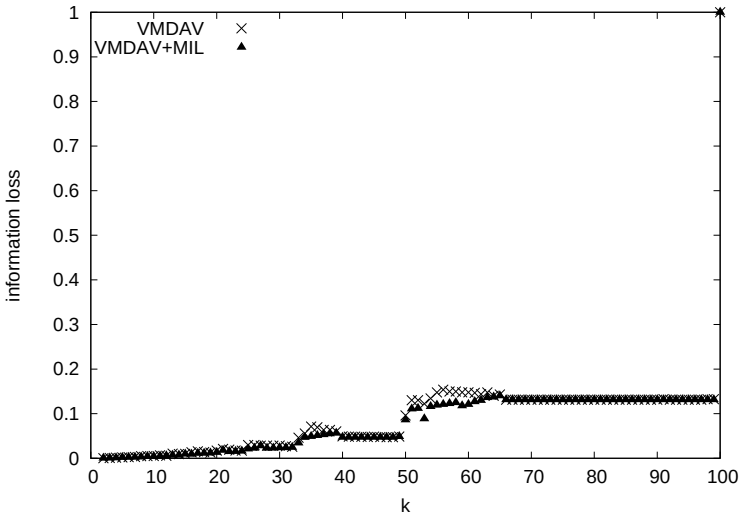

図 A.5: DS1 の情報損失の比較 (VMDAV)

表 A.3: DS1 の情報損失の値

k	MDAV	MDAV+MIL	VMDAV	VMDAV+MIL
2	0.000330	0.000330	0.000213	0.000205
3	0.000658	0.000533	0.000677	0.000546
4	0.001130	0.001130	0.001460	0.001021
5	0.001573	0.001573	0.001911	0.001808
6	0.002795	0.002154	0.002553	0.002119
7	0.002870	0.002441	0.003045	0.002417
8	0.003805	0.003805	0.004821	0.002930
9	0.004044	0.003882	0.004904	0.003923
10	0.004104	0.004104	0.005134	0.004041
11	0.005289	0.004850	0.006630	0.005012
12	0.009449	0.005724	0.005683	0.005608
13	0.010220	0.007043	0.010473	0.006713
14	0.007941	0.006907	0.009406	0.007138
15	0.012821	0.009420	0.009874	0.009407

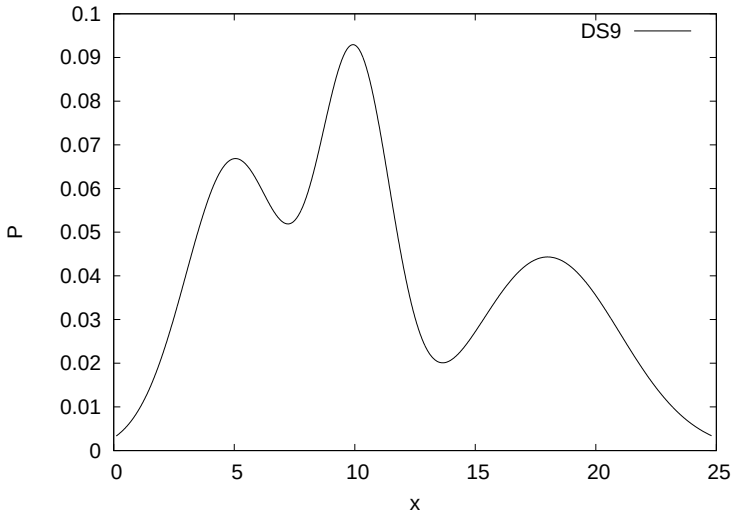

図 A.6: DS9 の確率分布

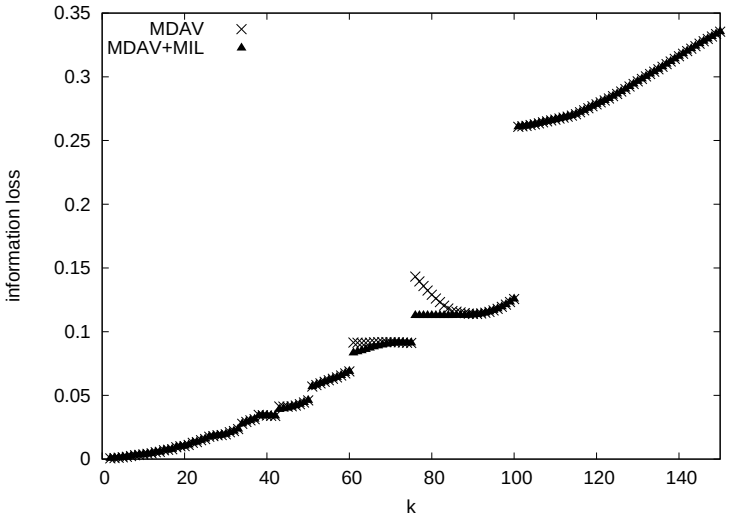

図 A.7: DS9 の情報損失の比較 (MDAV)

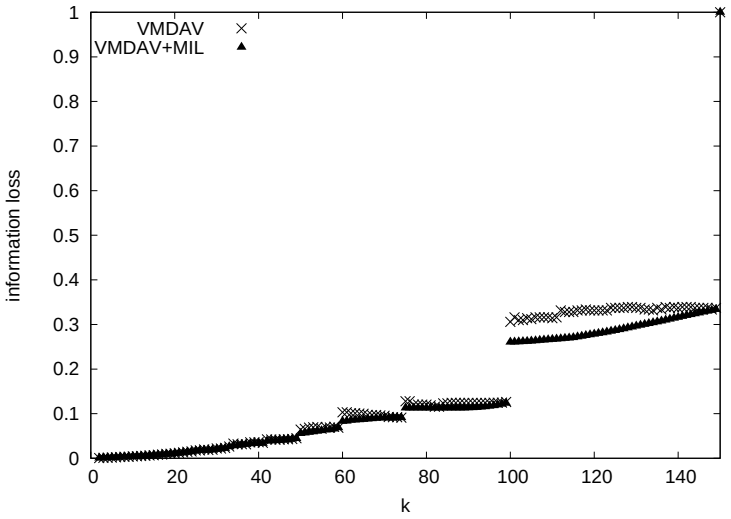

図 A.8: DS9 の情報損失の比較 (VMDAV)

表 A.4: DS9 の情報損失の値

k	MDAV	MDAV + MIL	VMDAV	VMDAV + MIL
2	0.000560	0.000560	0.000513	0.000490
3	0.000845	0.000845	0.000986	0.000848
4	0.001108	0.001108	0.001144	0.001109
5	0.001412	0.001412	0.001565	0.001444
6	0.001831	0.001831	0.002170	0.001857
7	0.002361	0.002361	0.002893	0.002635
8	0.003039	0.003035	0.003240	0.003043
9	0.003315	0.003311	0.003358	0.003125
10	0.003591	0.003591	0.004160	0.003694
11	0.004122	0.004120	0.004502	0.004096
12	0.004427	0.004427	0.005287	0.004539
13	0.005468	0.005467	0.005755	0.005533
14	0.005653	0.005632	0.006402	0.006175
15	0.006904	0.006904	0.006897	0.006745

表 A.5: MDAV に対する MDAV による情報損失の減少

データセット	減少した k の割合	減少率の平均値	減少率の最大値
DS0	69.4%	2.7%	7.6%
DS1	88.9%	35.5%	68.8%
DS2	87.9%	12.5%	31.2%
DS3	73.7%	11.9%	40.8%
DS4	80.8%	21.4%	54.9%
DS5	60.4%	10.4%	37.3%
DS6	55.0%	8.7%	40.0%
DS7	49.7%	14.6%	50.7%
DS8	73.2%	7.6 %	37.6%
DS9	38.9%	3.4%	21.3%
DS10	59.1%	18.5%	67.3%
DS11	81.9%	7.5%	32.2%
DS12	75.5%	8.6%	40.2%

表 A.6: VMDAV に対する MIL による情報損失の減少

データセット	減少した k の割合	減少率の平均値	減少率の最大値
DS0	91.8%	3.4%	21.7%
DS1	66.7%	14.3%	39.2%
DS2	70.7%	5.0%	19.8%
DS3	99.0%	7.5%	26.7%
DS4	97.0%	7.9%	35.4%
DS5	98.7%	11.5%	31.6%
DS6	91.3%	6.3%	34.1%
DS7	98.7%	16.8%	36.2%
DS8	98.0%	8.1%	32.7%
DS9	98.7%	8.9%	19.6%
DS10	82.6%	13.2%	51.7%
DS11	84.6%	4.5%	18.4%
DS12	69.4%	8.7%	27.6%

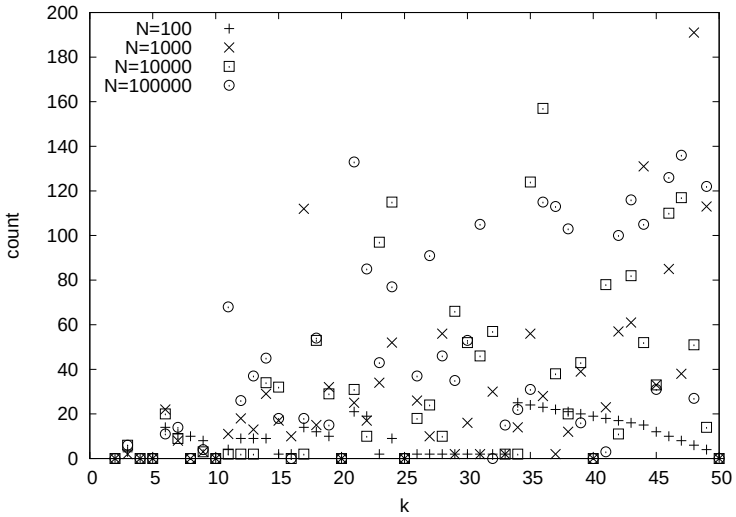

図 A.9: DS0 の確率密度関数に従うシード 0 のデータセットの判定回数 (MDAV)

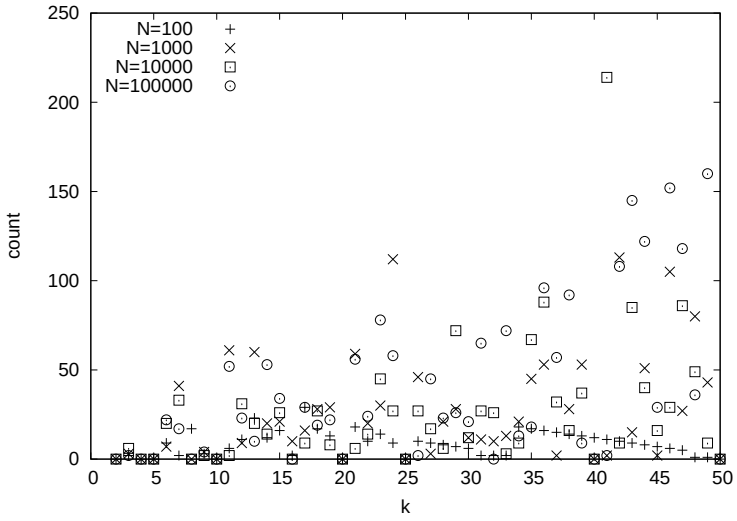

図 A.10: DS0 の確率密度関数に従うシード 1 のデータセットの判定回数 (MDAV)

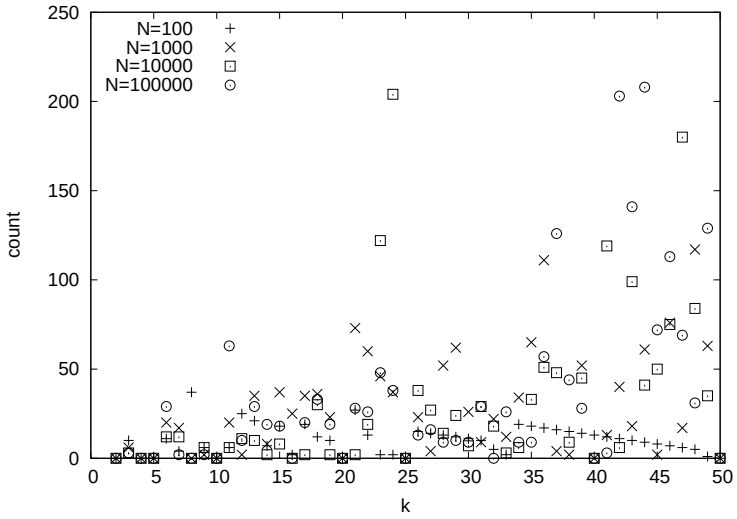

図 A.11: DS0 の確率密度関数に従うシード 2 のデータセットの判定回数 (MDAV)

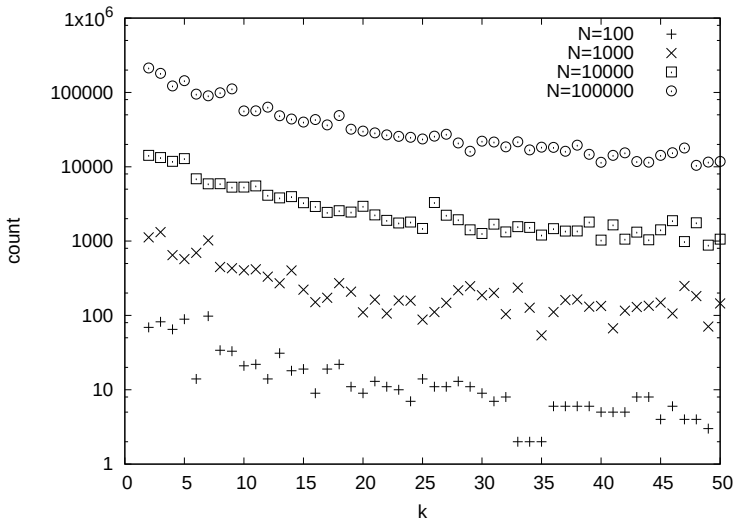

図 A.12: DS0 の確率密度関数に従うシード 0 のデータセットの判定回数 (VMDAV)

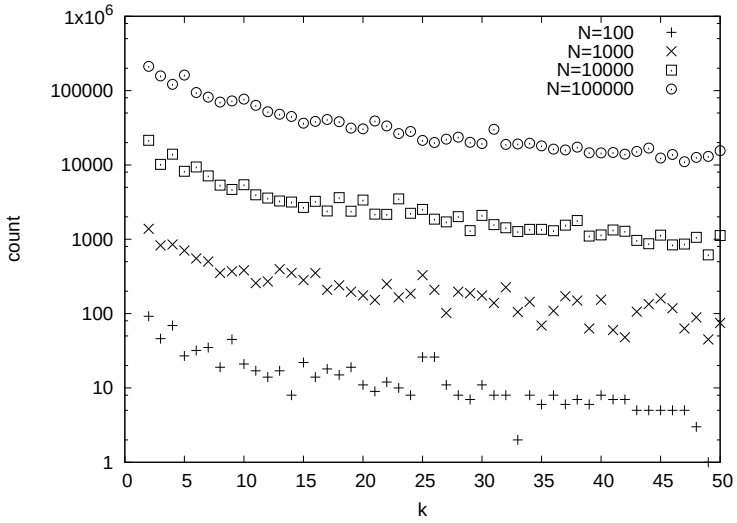

図 A.13: DS0 の確率密度関数に従うシード 1 のデータセットの判定回数 (VMDAV)

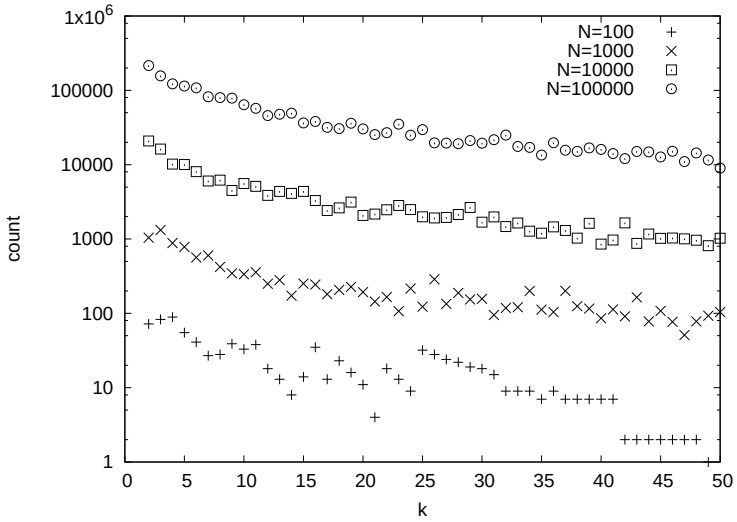

図 A.14: DS0 の確率密度関数に従うシード 2 のデータセットの判定回数 (VMDAV)

表 A.7: DS0 の確率密度関数に従うデータセットの判定回数の平均値

		N=100	N=1000	N=10000	N=100000
MDAV	seed0	9.06	29.53	33.76	44.94
	seed1	8.61	26.82	25.84	39.06
	seed2	10.10	27.90	30.45	35.53
VMDAV	seed0	18.08	277.16	3315.71	43016.80
	seed1	15.90	261.98	3327.88	41839.20
	seed2	19.41	256.63	3463.33	40274.00

表 A.8: DS0 の確率密度関数に従うデータセットの判定回数の最大値

		N=100	N=1000	N=10000	N=100000
MDAV	seed0	25	191	157	136
	seed1	27	113	214	160
	seed2	37	117	204	208
VMDAV	seed0	98	1319	14253	212956
	seed1	92	1381	21442	212065
	seed2	89	1322	20845	215233

表 A.9: データセットの判定回数の平均値と最大値

		N=100	N=1000	N=10000	N=100000
MDAV	平均値	9.23	33.58	29.05	34.21
	最大値	60	282	356	354
VMDAV	平均値	18.95	279.01	3384.47	41917.79
	最大値	151	1514	21960	249428