匿名化による情報損失に関する研究

（別紙１）

論　文　の　内　容　の　要　旨

氏　　名	秋山　寛子		
論文題目	（外国語の場合は、日本語で訳文を(　　)を付して記入すること。） 匿名化による情報損失に関する研究		
審査委員	区　分	職　名	氏　名
	委員長		印
	委　員		印
	委　員		印
	委　員		印
	委　員		印
	委　員		印

内　容　の　要　旨

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。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。

プライバシの保護のため、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守られる。しかしながらその一方で情報の持つ有用性は減少してしまう。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を見つけることが課題である。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尺度としては、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。

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。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、それをここでは情報容量とよぶ。データセットのデータの値を置き換えたときの情報容量の差の割合として、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り、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、これまでは適用が難しかった、数値と非数値の両方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というミクロアグリゲーションがある。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 の適用できるデータや応用先について述べる。

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。

（別紙２）

論 文 審 査 の 結 果 の 要 旨

<table>
<tr><td>氏　　名</td><td colspan="3">秋山　寛子</td></tr>
<tr><td>論文題目</td><td colspan="3">（外国語の場合は、日本語で訳文を(　　)を付して記入すること。）
匿名化による情報損失に関する研究</td></tr>
<tr><td rowspan="7">審査委員</td><td>区　分</td><td>職　名</td><td>氏　名</td></tr>
<tr><td>委員長</td><td></td><td>印</td></tr>
<tr><td>委　員</td><td></td><td>印</td></tr>
<tr><td>委　員</td><td></td><td>印</td></tr>
<tr><td>委　員</td><td></td><td>印</td></tr>
<tr><td>委　員</td><td></td><td>印</td></tr>
<tr><td>委　員</td><td></td><td>印</td></tr>
<tr><td colspan="4">要　　旨</td></tr>
<tr><td colspan="4">ビッグデータを活用して、新たな情報やサービスを提供することに対する関心が高まっている。ビッグデータには個人情報が含まれており、データの利用にあたっては不用意に個人情報が漏れないように十分注意を払う必要がある。このため、データから個人を特定・識別できないようにする加工する匿名化が行われる。個人情報を保護するためには情報の匿名化により匿名性を高めればよいが、その一方で匿名化によりデータの持つ情報の有用性が減少する。匿名化にはプライバシの保護と情報の有用性とのバランスをとることが必要である。そのためには、匿名化によって失われる情報量を測る指標が必要となる。
匿名化による情報の減少を測る目安として情報損失指標がいろいろ提案されている。しかし、これまでに定義されている情報損失指標は、特定のデータ種別や匿名化処理に対して定義されており、その活用は限定的なものとなっている。
本論文では、データの種別や匿名化処理に依存しない情報損失指標を考案し、その特性および応用について研究をおこなっている。まず情報容量と呼ぶ、データ間の距離に基づいたデータセット全体の持つ情報の量を定義している。匿名化処理を行った時、データセットの情報容量が減少する割合として、ILD（Information Loss based on Distance）という距離に基づく情報損失指標を独自に定義している。ILD 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を持っており、これまでは適用が難しかった数値と非数値の両方の属性を含むデータセットに対しても適用が可能である。
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た実験を行なっている。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というミクロアグリゲーションがある。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から、数値と非数値の両方の値に対して、どのようなグループ化を行うのが情報損失を最も小さくすることができるかを明らかにしている。</td></tr>
</table>

また ILD の応用として匿名化アルゴリズムの開発に応用可能であることを示している。ミクロアグリケーションによって k-匿名化を行う時、どのようにグループ分けを行えば情報損失を最小にすることができるか、最適なグループを決定することは非常に難しい問題である。本論文では、情報損失指標により匿名化アルゴリズムの構築が可能であることを示している。例えば、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している。このアルゴリズムでは、分割の修正を行うかどうかの判定を情報損失指標の定義から導出した式により行っている。このことから、このアルゴリズムにより、情報損失を極小にするグループ分けが得られることを示している。

以上のように、本論文はデータの種別に依存しないデータの情報容量および情報損失指標を構築することにより、数値データのみならず非数値データを含んだデータセットに対して、統一的に情報の有用性を測ることができることを示し、それを使って匿名化手法の違いによって情報損失がどのように変わるか比較することができることを示した。また、ミクロアグリケーションによる匿名化について、情報損失を極小にするグループ分けのアルゴリズムが構築できることを示している。これらの研究はビッグデータと呼ばれる様々なデータを活用する上で意義のある研究である。この論文の内容についてはすでに情報処理学会論文誌 Vo. 57, No.12 および情報処理学会論文誌「数理モデル化と応用」Vol. 11, No.3 に掲載され、公表されている。

よって、本学位申請論文は、奈良女子大学博士（情報科学）の学位を授与されるに十分な内容を有していると判断した。