2E-5

比較可能な匿名化グループを生成する匿名化手法の提案

豊田　由起[†]　　宮川　伸也[†]　　側高　幸治[†]　　伊東　直子[†]

日本電気株式会社　サービスプラットフォーム研究所[†]

1. はじめに

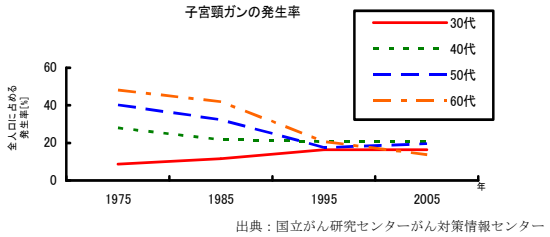
企業や病院等によって収集されたユーザの情報をより活用するための 1 つの方法として、第三者へ公開することで二次活用することが考えられる。公開する情報に、病歴や病状のセンシティブな情報が含まれる場合、個人のプライバシに留意しなければならない。

プライバシを保護するための手法として匿名化がある。従来の匿名化では、複数のユーザの情報が含まれるデータセットに対して、内容が近い情報同士をグループ化し匿名化したときに、手法によっては匿名化グループ同士のデータ数の比較を行えない。また、データの増加に伴って匿名化を行う度に、異なる匿名化グループが形成される場合があり、匿名化グループのデータの変化を追えない課題がある。本稿では、増加するデータに対して、匿名化グループに属するデータ数の構成内容に着目し、データ数を比較可能な匿名化グループを生成するアルゴリズムを提案する。

2. 特異点集合追跡の重要性

複数のユーザ情報をグループ化したときに、要素数がある基準値よりも少ないグループを、ここでは特異点集合と呼ぶ。

時間の経過に伴って増加するデータセットに対して、逐次グループ化した場合、特異点集合であったグループが、ある時点から特異点集合ではなくなる場合がある。例えば、10%を特異点集合の判断基準としたとき、図 1 に示すように、子宮頸ガン患者の 30 代の発生率は、1975 年頃は特異点集合であった。しかし、1985 年頃には特異点集合ではなくなっており、その原因を調査し、発生率を減少させる対策を検討する必要がある。このことは、特異点集合に着目してその動向を追跡することの重要性を示している。



出典：国立がん研究センターがん対策情報センター

図 1　子宮頸ガンの発生率

3. 従来の匿名化の課題

k-匿名化[1]は、情報の抽象化等によって k 人以上のグループを形成する処理である。本稿では、この指標を k と表記する。この処理では、特異点集合が形成される場合がある。情報の抽象化手法の一つであるグローバルリコーディング[2]では、特異点集合は切り落とされるかまたは、特異点集合とそれ以外の特定のグループ全体が抽象化される。ローカルリコーディング[3]と呼ばれる手法では、特定のグループから抽象化される情報を一部に制限することで情報の損失を抑制している。しかし、[3]の手法には次のような課題がある。

● 課題 1：ある時刻でグループの要素数を比較できない

従来の手法では、ある時刻 t_0 において特異点集合 A を匿名化する場合、k を満たすために、不足する数のみ他の非特異点集合 B から要素を取り出して、特異点集合 A と共に抽象化する。非特異点集合 B 以外の非特異点集合 C が存在する場合、非特異点集合 B と C の要素数の比較結果が匿名化前後で異なってしまう。

● 課題 2：時系列でグループの変化を追えない

時刻 t_0 から時間 t が経過した時刻 t_1 において再び特異点集合 A を匿名化する場合、非特異点集合 B とは異なる非特異点集合 D から要素を取り出す可能性がある。仮に、非特異点集合 B から取り出したとしても、特異点集合 A の要素数が増加すれば非特異点集合 B から取り出す要素が減る。そのため、時刻 t_0 と t_1 の非特異点集合 B と D の要素数の比較結果が匿名化前後で異なってしまう。(課題 2-1)

また、時刻 t_1 において、新たに特異点集合 E が出現する可能性もあり、その場合に非特異点集合 B や D から取り出す要素数が増え、非特異点集合 B や D の要素数の比較結果が匿名化前後で異なってしまう。(課題 2-2)

4. 提案アルゴリズム

本稿で提案する匿名化アルゴリズムは、(1)非特異点集合から特異点集合に含めて抽象化する抽象化要素数を計算し、(2)抽象化要素を各グループから取得して抽象化する。

4.1 抽象化要素数の計算

各グループの抽象化要素数(以下、p)を決定す

るアルゴリズムを述べる。各グループは抽象度に基づいて図 2のように木構造をなしているとする。まず、トップノードに対して、k をノードの子ノードの数で割り、各子ノードの p を求める。次に、各子ノードに対して、それぞれ、k とノードの p を加算した値を、ノードの子ノード(孫ノード)の数で割り、各孫ノードの p を求める。子ノードと同様の処理を木構造の最下層まで繰り返す。

同じ親ノードを持つ全ての非特異点集合から同じ要素数が取り出されるように p を計算することにより、非特異点集合の要素数の差が匿名化前後で同じになるため、課題 1 を解決できる。また、各ノードが k 以上の要素数を持つように抽象化要素数を決定することにより、時刻 t_1 においていずれかのノードの下に新たな特異点集合が出現しても、特異点集合を一段階抽象化するのみで済む。このとき、他の非特異点集合から取り出す要素数が変わらないため、課題 2-2 を解決できる。

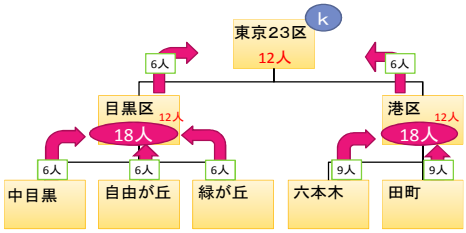


図 2　抽象化要素数の算出例

図 2 は p の算出例である。自宅最寄り駅である「中目黒」「自由が丘」「緑が丘」と、それらをさらに抽象化した概念として、区名の「目黒区」、さらに、区名の「目黒区」「港区」を抽象化した概念として「東京 23 区」が定義されているとする。以下の例では k=12 とする。トップノードである「東京 23 区」の子ノードの数は「目黒区」「港区」の 2 つであるため、k(12)と子ノードの数(2)から、子ノードの p は 12/2=6 となる。同様に「目黒区」の子ノードの p は、(12+6)/3=6 となる。

4.2 要素の抽象化

要素を抽象化するアルゴリズムを述べる。まず、最下層のグループから順に「グループの要素数」と「上記で求めた p と k の値の和」を比較し、グループの要素数の方が大きければ、グループから p 個の要素を取り出して抽象化し、その他の場合は、グループの全要素を取り出して抽象化する。二回目以降は、各ノードの p の値を初回と同じ値にする。その結果、各非特異点集合からの p が一定になるため、課題 2-1 と課題 2-2 を解決できる。

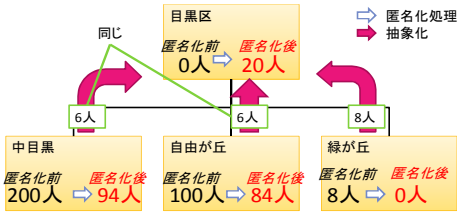


図 3　抽象化の具体例

図 3 に抽象化の具体例を示す。匿名化前の「中目黒」「自由が丘」「緑が丘」に所属する人数を、200 人、100 人及び 8 人とする。

「中目黒」「自由が丘」「緑ヶ丘」は、それぞれ前述の通り p=6 である。各グループから抽象化要素を取り出しても非特異点集合であるための基準値である(k+p=12+6=)18 人以上の人数を含む「中目黒」と「自由が丘」の 6 人を「目黒区」に抽象化する。一方、「緑ヶ丘」の人数は 18 人を満たさないため、8 人全員を「目黒区」に抽象化する。

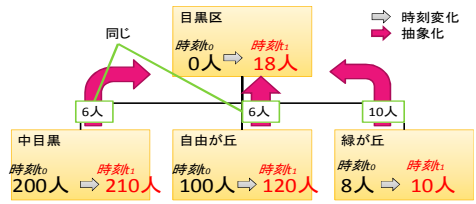


図 4　時刻 t_1 における抽象化の具体例

図 4 では、時刻 t_1 において「中目黒」「自由が丘」「緑が丘」の人数が、210 人、120 人及び 10 人に増加したとする。それぞれ p=6 であるため「中目黒」「自由が丘」から「目黒区」に 6 人ずつ抽象化し、「緑が丘」に所属する人数は、(k=)12 人を満たさないため、10 人全員を抽象化する。

5. おわりに

本稿では、データ数を比較可能な匿名化グループを生成するアルゴリズムを提案し、その具体例を説明した。今後は、実データに対して本手法を適用して、その有効性を示すことを目指す。

本研究は、総務省「平成 22 年度大規模仮想化サーバ環境における情報セキュリティ対策技術の研究開発」の一環として実施している。

参考文献

[1]P. Samarati, *Protecting Respondents' Identities in Microdata Release*, IEEE Trans. on Knowl. and Data Eng. 13(6), pp. 1010-1027, 2001.
[2]K. LeFevre, *Incognito:Efficient Full-Domain K-Anonymity*, ACM SIGMOD Int'l Conf. on Management of Data, pp. 49-60, 2005
[3]J. Xu, *Utility-Based Anonymization Using Local Recoding*, ACM SIGKDD Int'l Conf. on Kowledge discovery and datamining, pp. 785-790, 2006