属性値のグループ分類を用いたPk-匿名化手法の検討

柿澤 美穂 †　　渡辺 知恵美 ††　　古川 諒 †††　　高橋 翼 †††

† お茶の水女子大学
112-8610 東京都文京区大塚 2-1-1
kakizawa.miho@is.ocha.ac.jp

†† 筑波大学
305-8573 茨城県つくば市天王台 1-1-1
chiemi@cs.tsukuba.ac.jp

††† 日本電気株式会社
211-8666 神奈川県川崎市中原区下沼部 1753
r-furukawa@cb.jp.nec.com, t-takahashi@nk.jp.nec.com

あらまし　ランダム化を用いた匿名化手法の一つに Pk-匿名化がある. 既存の Pk-匿名化は, データ主体を 1/k 以上の確信度に絞り込めないよう, 元の属性値にラプラス分布に従うノイズを付与することで実現されている. 我々は先行研究にて, 既存手法におけるノイズが過剰に付与される点の解決策として, 元の属性値を予め複数のグループに分類してから Pk-匿名化を実現する手法を提案し, 属性値を予め複数のグループに分類した場合, 既存手法と比較してラプラス分布の分散をより小さく抑えられることを示した. 本稿では, ラプラス分布の分散を抑えて Pk-匿名化を実現するのに有効な属性値のグループ分類方法として, Mondrian と DBSCAN による濃度ベースクラスタリングを用いた手法を提案し有効性を検証する.

Improvement of Pk-anonymization using grouping of attribute values

Miho Kakizawa†　　Chiemi Watanabe††　　Ryo Furukawa†††
Tsubasa Takahashi†††

†Ochanomizu University
2-1-1 Otsuka,Bunkyo,Tokyo 112–0012, JAPAN
kakizawa.miho@is.ocha.ac.jp

††University of Tsukuba
1-1-1 Tennodai,Tsukuba,Ibaraki 305–8573, JAPAN
chiemi@cs.tsukuba.ac.jp

†††NEC Corporation
Kawasaki,Kanagawa 211-8666, JAPAN
r-furukawa@cb.jp.nec.com, t-takahashi@nk.jp.nec.com

Abstract　Pk-anonymization is a data anonymization method that employs randomization. To implement existing method, we assign random noise using a Laplace distribution to records to reduce the probability of identifying record owners to less than 1/k. In our study, as a solution of assigning noise excessively, we proposed the method of implementation of Pk-anonymization with grouping original attribute values in advance. In this case, we can implement Pk-anonymization with smaller variance. In this paper, as a grouping method, we propose a method using Mondrian and density-based clustering with DBSCAN to achieve Pk-anonymization by suppressing the variance of the Laplace distribution and verify the validity.

1 はじめに

近年, データベースサービスの普及に伴い, 個人情報等の機密情報をデータベースに格納する際にプライバシ保護が要求されている. 特に, データベースに格納された機密情報を公開する際, データ公開者はデータベースのレコード所持者をデータ利用者に特定させないよう, レコード所持者を隠す匿名化手法が研究されている. データ匿名化の一手法として, k-匿名化 [10] [6] がある. k-匿名化とは, 属性値の抽象化, 削除等を行うことによって, レコード所持者を k 人未満に絞れないようにする手法である. この k-匿名化を確率的指標に拡張した手法として, Pk-匿名化 [2] [3] [4] がある. Pk-匿名化は, 属性値の置換やノイズ付与といった確率的な操作を用いて, レコード所持者を 1/k 以上の確信度で絞り込めないようにする. 既存研究では, 数値属性に対してラプラス分布に従うノイズを付与することで Pk-匿名化を実現する方法が提案されている. 既存の Pk-匿名化では, ラプラス分布の分散値を決定する際に, 属性値間の最大距離とレコード数を用いる. しかし, 属性値の分布によっては, この方法では過剰にノイズが付与され匿名化が過剰になされる場合がある. 例えば, レコードの属性値の分布に特徴があり, いくつかのグループに明確に分類できるような場合である. このようなデータに対して Pk-匿名化を適用する際, レコード全体に対してラプラス分布に従うノイズを適用すると, ラプラス分布の分散が大きくなり匿名化前に見られていたデータの特徴が失われてしまう場合がある. そこで我々は, ラプラス分布に基づくノイズ適用による Pk-匿名化手法において, 匿名化前のデータの特徴をできるだけ維持するための改良法を文献 [1] にて提案した. 匿名化前のデータの類似性を考慮していくつかのグループに分類し, 各グループの属性値集合に対して Pk-匿名化を適用する. 属性値を予めグループに分類してから Pk-匿名化を施すことで, 各グループに Pk-匿名化を適用した際のラプラス分布に従うノイズの分散は, 既存手法に比べて小さく抑えられることを示した. そこで本稿では, 属性値をグループに分類する手法として, DBSCAN[8] による濃度ベースクラスタリングと Mondrian[7] を用いて, ラプラス分布に従うノイズの分散を小さく抑えるために有効な手法を提案する.

2 グループ分類を用いた Pk-匿名化

文献 [1] にて, 我々は属性値をいくつかのグループに分類してから既存手法の Pk-匿名化を施すことで, 属性値全体に Pk-匿名化を施した場合よりもラプラス分布に従うノイズの広がりを抑えられることを示した. ノイズが従うラプラス分布の分散は, 属性値間の最大距離とレコード数に基づいて決定されている. 属性値をグループ分類することで, 属性値間の最大距離を小さくすることができ, したがってラプラス分布の分散も抑えることができる. そこで我々は, ラプラス分布の分散を抑えるための属性値のグループ分類方法を二つ提案する.

手法 1 : Mondrian を用いた分類法

手法 2 : Mondrian と DBSCAN を併用した分類法

以下, 二つの方法について, 各方法を用いてグループに分割し, グループ毎に Pk-匿名化を施した場合の, ノイズが従うラプラス分布の分散値の比較結果について述べる.

3 Mondrian と DBSCAN

属性値をグループに分類する手法として, Mondrian[7] と DBSCAN[8] を用いた方法を提案する. 本節では, この二つの既存手法を説明する.

3.1 手法 1: Mondrian を用いた分類法

Mondrian では, レコードの集合に対して分割する次元を決め, その次元の中央値に基づいて各レコードを左右に分類していく. この操作を, レコードの集合がこれ以上分割されない状態になるまで再帰的に繰り返す. Mondrian を用いると, 近くに分布する属性値が同じグループに

属する結果になり, 属性値間の最大距離も自然と小さくなる. 結果として, ラプラス分布に従うノイズの分散も小さくなることが期待される.

手法 1 では, データ全体に対し Mondrian を適用し, 属性値をいくつかのグループに分類する.

3.2 手法 2: Mondrian と DBSCAN を併用した分類法

DBSCAN とは, ある一点を基点とし, 同じクラスタに含めることができる条件を満たす点を推移的にたどっていき, 到達可能な点の極大集合を一つのクラスタとするクラスタリングの一手法である. DBSCAN を用いると, 大まかないくつかのクラスタに分類でき, クラスタから大きく離れて分布している属性値を外れ値としてグループから除外することができる. そのため, 外れ値として扱われる属性値を含めたグループを Pk-匿名化するよりも, 属性値の最大距離を抑えることができ, 結果としてラプラス分布の分散も小さく抑えることができる.

手法 2 では, 初めに DBSCAN をデータ全体に対して適用し, その後 DBSCAN によって分割されたクラスタ毎に Mondrian を適用する. DBSCAN を適用すると, 外れ値となる属性値はクラスタに含まれない. 我々は, この外れ値となる属性値をノイズとし, DBSCAN を適用することでノイズを除去して, 属性値間の最大距離を小さくすることを目的とする. DBSCAN を適用しただけでは, ノイズが除かれた大まかなクラスタに分類されるだけであり, そのクラスタ内の属性値間の最大距離はデータ全体の属性値間の距離と大きく変わらない. そのため, DBSCAN を適用した後のデータに Mondrian を適用することで, ノイズを除去した状態からレコード数の小さいグループに分類することができ, 結果として属性値間の最大距離が小さいグループを複数作ることができる. この時, DBSCAN で分類した大まかなクラスタ毎に Mondrian を適用することで, クラスタを跨いだグループを形成しないようにしている. クラスタを跨いでグループを形成してしまうと, 属性値間の最大距離が大きくなってしまうからである.

このように, 手法 1 と手法 2 の二つの属性値グループ分類手法を用いて, サンプルデータに対して属性値をグループに分類して Pk-匿名化を施した場合の, ラプラス分布の分散を比較する.

4 グループ分類を用いた検証実験

これまでに述べてきた Mondrian と DBSCAN を用いたグループ分類方法を実際にサンプルデータに適用し, Pk-匿名化を施した場合の結果を比較する. ここで使用するサンプルデータとして, 三種類のデータセットを用意した.

1. 大まかな二つのまとまりを形成しているデータ (相関なし)(dataA)
2. 大まかな二つのまとまりを形成しているデータ (相関あり)(dataB)
3. 大まかな五つのまとまりを形成しているデータ (dataC)

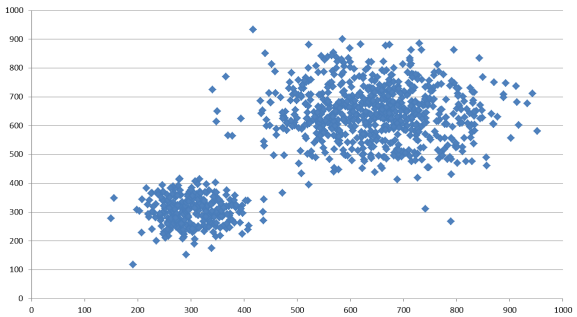

図 1: dataA

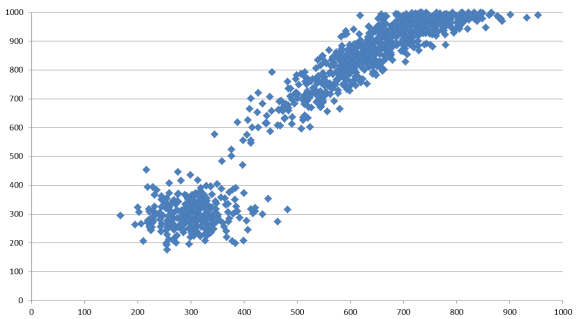

図 2: dataB

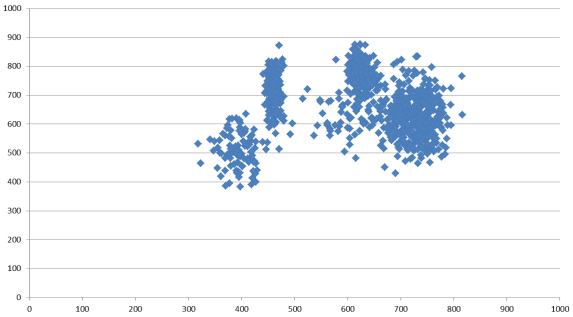

図 3: dataC

Mondrian を適用する際には, グループに含める属性値の最小数を指定すると, その数より大きい数の属性値が含まれたグループ分類が行われる. 本研究での Mondrian は, 属性値がグループ間で重複することなく分類するよう設定している. 本実験では, Mondrian によって属性値の最小数を 100 として属性値をグループに分類し, 形成されたグループに対し k=5 として Pk-匿名化を施す. 三つのサンプルデータに対し, 上記の条件で実験した結果を以下に述べる.

4.1 dataA に対する実験結果

まず, dataA に対し手法 1 で Mondrian を適用した場合, 以下の図 4 のようなグループ分類結果を得る.

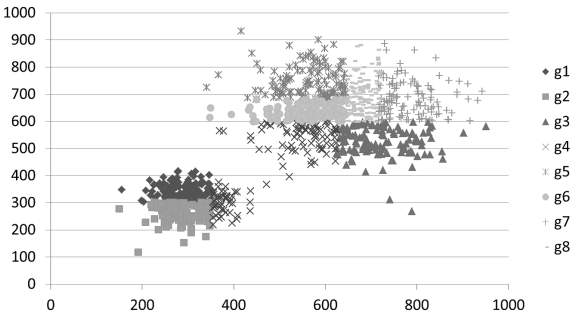

図 4: dataA に対し Mondrian を適用した場合

120〜130 個の属性値を含むグループが 8 個形成された. この 8 グループに対し, 各グループ毎に Pk-匿名化した場合の分散値を, グループ分類せずデータ全体に対して Pk-匿名化した場合の分散値で割った値を比較すると, 以下の図 5 のようになる.

	g1	g2	g3	g4	g5	g6	g7	g8	average
x/all	0.395559	0.404737	0.657435	0.542147	0.614103	0.591798	0.457278	0.140688	0.475468
y/all	0.223931	0.360253	0.650309	0.749901	0.50017	0.160365	0.564691	0.552816	0.470305

図 5: dataA の手法 1 適用時の分散値比較

上段の値は, x 軸の値域に対するラプラス分布の分散値で, 下段の値は, y 軸の値域に対するラプラス分布の分散値で計算した結果である.

一方, 手法 2 で DBSCAN を適用してから, DBSCAN によって分類されたクラスタ毎に Mondrian を適用した場合のグループ分類結果は, 以下の図 6 のようになる.

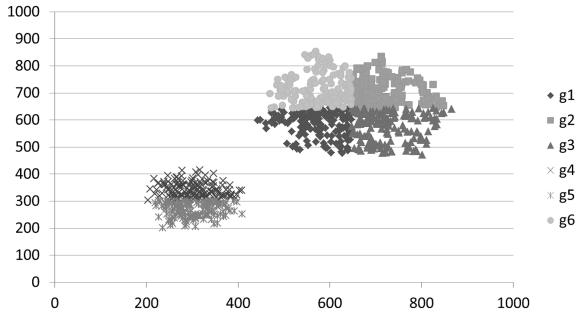

図 6: dataA に対し DBSCAN と Mondrian を併用した場合

140〜160 個のグループが 6 個形成され, ノイズとなる属性値が除外されている. この 6 グループに対して Pk-匿名化を施し, 同様の計算をした結果が以下の図 7 である.

	g1	g2	g3	g4	g5	g6	average
x/all	0.433782	0.46882	0.400066	0.456908	0.444174	0.488981	0.448788
y/all	0.236421	0.26274	0.473823	0.382192	0.435303	0.392623	0.36385

図 7: dataA の手法 2 適用時の分散値比較

手法 1 と手法 2 の結果の平均値を比較してみると, 手法 2 の場合の方がラプラス分布のノイズの分散を小さく抑えられていることが分かる. これと同様の実験を, 他の二つのサンプルデータに対して行った結果を以下に示す.

4.2 dataB に対する実験結果

まず, 手法 1 を適用した場合のグループ分類結果は以下の図 8 のようになる.

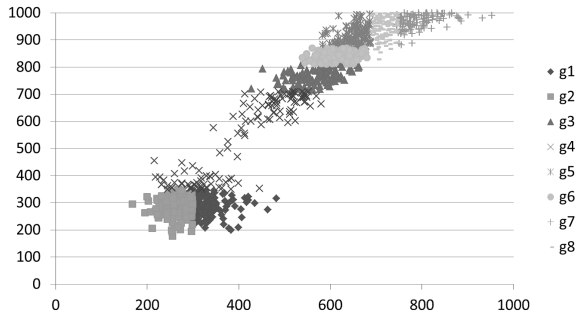

図 8: dataB に対し Mondrian を適用した場合

120〜130 個の属性値を含む 8 グループが形成された．この各グループに対して Pk-匿名化を施した場合に, 全体データの分散値に対するグループの分散値の割合を比較すると, 以下の図 9 のようになる．

	g1	g2	g3	g4	g5	g6	g7	g8	average
x/all	0.374592	0.266953	0.485729	0.741066	0.21358	0.286395	0.431289	0.115499	0.364388
y/all	0.291279	0.333481	0.193153	0.699704	0.240412	0.114384	0.234409	0.333694	0.305064

図 9: dataB の手法 1 適用時の分散値比較

一方, 手法 2 を適用した場合のグループ分類結果は以下の図 10 のようになる．

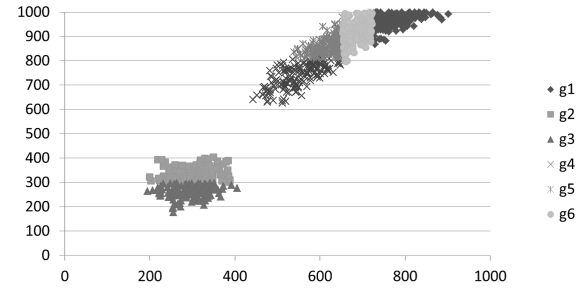

図 10: dataB に対し DBSCAN と Mondrian を併用した場合

130〜170 個の属性値を含む 6 グループが形成された．全体データの分散値に対する, これらの各グループの分散値の割合は, 以下の図 11 である．

	g1	g2	g3	g4	g5	g6	average
x/all	0.409971	0.457821	0.239501	0.425978	0.372343	0.143077	0.341448
y/all	0.196617	0.228052	0.300823	0.326567	0.237828	0.359222	0.274851

図 11: dataB の手法 2 適用時の分散値比較

手法 1 と手法 2 の結果の平均値を比較してみると, dataA に対する実験の時と同様に, 手法 2 の場合の方がラプラス分布のノイズの分散を小さく抑えられていることが分かる．

4.3 dataC に対する実験結果

大まかな五つのクラスタに予め分かれている dataC に対して, 手法 1 を適用した場合のグループ分類結果は以下の図 12 のようになる．

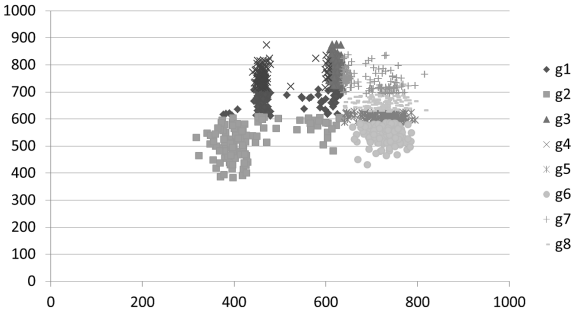

図 12: dataC に対し Mondrian を適用した場合

他と同様,120〜130 個の属性値を含む 8 グループが形成された．この各グループに対して Pk-匿名化を施した場合に, 全体データの分散値に対するグループの分散値の割合を比較すると, 以下の図 13 のようになる．

	g1	g2	g3	g4	g5	g6	g7	g8	average
x/all	0.830343	1.013184	0.094704	0.534266	0.499164	0.411787	0.577862	0.571551	0.566608
y/all	0.335016	0.730145	0.52112	0.519996	0.132607	0.502238	0.473816	0.202043	0.427123

図 13: dataC の手法 1 適用時の分散値比較

一方, 手法 2 を適用した場合のグループ分類結果は以下の図 14 のようになる．

80〜100 個の属性値を含む 11 グループが形成された．全体データの分散値に対する, これらの各グループの分散値の割合は, 以下の図 15 である．

手法 1 と手法 2 の結果の平均値を比較してみると, 他のサンプルデータに対する実験の時と同様に, 手法 2 の場合の方がラプラス分布のノイズの分散を小さく抑えられていることが分かる．

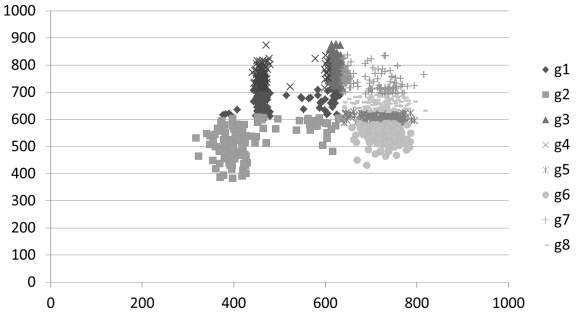

図 14: dataC に対し DBSCAN と Mondrian を併用した場合

	g1	g2	g3	g4	g5	g6	g7	g8	g9	g10	g11	average
x/all	0.550347	0.542171	0.274932	0.236904	0.154736	0.359185	0.628691	0.622504	0.420632	0.148517	0.196504	0.421122
y/all	0.202117	0.533945	0.255629	0.516421	0.417084	0.399215	0.486843	0.197942	0.840227	0.364694	0.502952	0.427714

図 15: dataC の手法 2 適用時の分散値比較

4.4 グループ分類方法に関する考察

これら三つのサンプルデータセットに対する実験結果において, どの場合も手法 2 の DBSCAN を適用してから, DBSCAN によって分類されたクラスタ毎に Mondrian を適用する場合の方が, ノイズが従うラプラス分布の分散を小さく抑えられていることが分かった. DBSCAN でまずノイズとなる属性値を除去した大まかなクラスタを形成し, そのクラスタ毎に Mondrian を適用して属性値をグループに分類したことで, 属性値間の最大距離を二段階にわたって小さくすることができたため, グループ内の属性値間の最大距離も小さくなり, ラプラス分布の分散も抑えることができたと考えられる. したがって, 我々の提案する二つの手法においては, 手法 2 の DBSCAN と Mondrian を併用する手法が有効であることが示された.

5 まとめと今後の課題

我々は, 機密情報を公開する際にデータを匿名化する手法の一つである Pk-匿名化の改良法として, 属性値をあらかじめいくつかのグループに分類してから, グループ毎に Pk-匿名化することで, ノイズの過剰付与を防ぐ手法を提案した. 本稿では, 属性値のグループ分類方法に焦点を置き, Mondrian と濃度ベースクラスタリング手法である DBSCAN を取り入れた手法を提案し, 比較, 評価を行った. DBSCAN を適用してノイズとなる属性値を除去してから, DBSCAN によって分類されたクラスタ毎に Mondrian を適用して属性値をいくつかのグループに分類することで, 匿名化後の値に付与される, ラプラス分布に従うノイズの分散を小さく抑えられることを示した. 今後の課題としては, 密度を考慮したグループ分類方法の考案や, 匿名化後の値の有用性の評価などが挙げられる.

参考文献

[1] 柿澤 美穂, 渡辺 知恵美, 古川 諒, 高橋 翼,”Pk-匿名化手法の一改良法の検討”, DEIM2014, 2014

[2] 五十嵐 大, 千田 浩司, 高橋 克巳,”k-匿名化の確率的指標への拡張とその適用例”, CSS2009, 2009

[3] 五十嵐 大, 千田 浩司, 高橋 克巳,”数値属性における k 匿名化を満たすランダム化手法”, CSS2011, 2011

[4] 五十嵐 大, 千田 浩司, 高橋 克巳,”ランダム化データベース上の k-匿名性の一般的算出法”, CSS2011, 2011

[5] 五十嵐 大, 長谷川 聡, 納 竜也, 菊池 亮, 千田 浩司, 数値属性に適用可能なランダム化により k 匿名化を保証するプライバシクロス集計, CSS2012, 2012

[6] 千田 浩司, 木村 映善, 五十嵐 大, 濱田 浩気, 菊池 亮, 石原 謙,”集合匿名化データの多変量解析評価”, CSS2012, 2012

[7] Kristen LeFevre, David J. DeWitt, Raghu Ramakrishnan, ”Mondrian Multidimensional K-Anonymity,” ICDE, 2006

[8] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, ”A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, KDD-96, 1996

[9] J. Li, "Preservation of proximity privacy in publishing numerical sensitive data", SIGMOD2008, 2008

[10] L. Sweeney. "k-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,10(5), pp.555-570, 2002.

[11] R. Agrawal and R. Srikant. Privacy-preserving data mining. Proc. of the 2000 ACM SIGMOD Intl. Conf. on Management of Data, 2000