頻度情報の付加による匿名化データの有用性向上技術の一提案

寺田剛陽[1]　山岡裕司[1]　福岡尊[1]

概要： AI を活用した高精度の分析には多くのデータを要する．自組織のデータだけで充分な精度が得られない場合はオープンデータや他組織のデータの活用が考えられるが，これらのデータは個人情報漏洩の防止の観点から匿名化が施されている．匿名化はデータの情報量を減らしてしまうため AI での分析精度に影響する．本論文では，匿名化によるデータ劣化を抑制するため，匿名化前の統計情報を匿名化後のデータに付与する方式を提案し評価した．結果，ロジスティック回帰，線形サポートベクター分類器との相性がよい傾向がわかった．

A Proposal of Utility Improvement of Anonymized Data by Adding Frequency Distribution

TAKEAKI TERADA[1]　YUJI YAMAOKA[1]　TAKERU FUKUOKA[1]

1. はじめに

昨今，機械学習技術の進歩により，画像識別や事象分類の精度が飛躍的に向上した．機械学習の精度は，入力とするデータの量が多いほどより高くなるが，そのためには一企業が所有するデータだけでなく，他の組織が提供するデータも利用することが必要である．しかしながらデータには個人情報が含まれている場合が多く，そのままでは提供されない．結果的に他の組織から調達されるデータは個人情報が特定されないように加工されたデータが一般的である．

この加工技術として，k-匿名化[1]がある．これが想定するデータは個票データとよばれるレコードの集合である．1 件のレコードは 1 人のユーザのデータに対応していて，レコードはたとえば年齢，性別，職業など 1 つ以上の属性で構成され，各レコードはそれぞれの属性について値をもつ．k-匿名化は，属性値の組合せが同じレコードが k 個以上になるようにレコード集合をグルーピングし，属性値の一般化によりそのグループ内で属性値の組合せを同じにする技術であり，個人情報の特定を防ぐ有用な手段であるが問題もある．それはデータが含む情報の量が落ちることである．匿名化処理の際に，各レコードを特定のユーザのデータであると断定できないよう，同じ属性値の組合せ（等価クラスとよばれる）を持つレコードを探索するが，見つからない場合は各レコードの一部の属性の値を削除するか，それらのレコードの属性値をより一般化した同一の値に置き換えることで等価クラスを作る．この結果，データの情報量が下がるため，このデータで学習した予測モデルは，匿名化する前のデータを使って学習した予測モデルよりも精度が落ちてしまう傾向がある．

精度低下を抑える従来技術として，k-匿名化によりグルーピングされたレコード集合ごとに，匿名化前のデータの平均値や分散，確率質量関数といった統計情報を付加する手法[2]（従来手法 1 とよぶことにする）がある．しかしながらこの手法は，匿名化データ提供事業者が匿名化データと匿名化前データの照合が容易にできる（提供元基準で容易照合性がある）場合があるため，この手法により作成されたデータは日本では個人情報保護法により第三者提供ができない可能性がある．提供元基準とは，日本の「個人情報の保護に関する法律についてのガイドライン（通則編）」において，データを第三者提供する際に当該データが個人情報か否かの判断基準として一般的に用いられている考え方である．そこで我々は，統計情報をグループ単位ではなくデータセット単位について算出し，匿名データに付与することで提供元基準での容易照合性のない手法を提案する．実験の結果，従来手法 1 が比較対象とした，統計情報を付加しない手法（従来手法 2 とする）よりも分類精度が 0～5%向上した．

本論文の貢献は次の通りである．

- 匿名化データを機械学習の入力データとして活用する際の有用性を高める手法として Inan ら[2]の方式があるが，この安全性を改善するデータエンコーディング方式を提案し，オープンデータでの精度検証の結果，Inan らが比較対象とした従来方式よりも精度が向上したことを確認した．

以降，本論文は次のように構成される．2 章では k-匿名

1 株式会社 富士通研究所
FUJITSU LABORATORIES LTD.

化の従来研究を紹介し，3 章では従来研究のうち提案方式がベースとした方式の課題を述べる．4 章で提案方式，5 章で実験による評価結果を示し，6 章で考察，7 章でまとめる．

2. 従来技術

k-匿名化は 2002 年に Sweeney ら[1]が最初に提案した匿名化技術であり，個票データを対象とする．個票データとは個人を単位とした情報（レコードと呼ばれる）の集合データである．各レコードは特定の個人のデータに対応する．個票データは 1 つ以上の属性についてのデータであり，各レコードは属性ごとに個別の値を持つ．属性は一般的には識別子，準識別子（QIs : Quasi-Identifiers）とセンシティブ属性に分類される．識別子はデータの持ち主を個人に特定できる属性であり，個人 ID などが挙げられる．匿名化の際，識別子は削除される．準識別子とは，単体では個人を識別できないが，複数を組み合わせることで個人を識別できるようになる属性であり，年齢や性別，職業，居住地などが挙げられる．k-匿名化では，個票データから任意のレコードを選んだ時に，各列のデータのうち準識別子である列のデータの組み合わせが全く同じものが最低でも k 個以上あるように個票データを加工する．センシティブ属性は機微情報を表す属性であり，たとえば基礎疾患（高血圧，糖尿病，がんなど），年収などが挙げられる．k-匿名化は，適用されたデータセットにおける個人情報を守る効果が得られる一方で，最低 k レコード以上の同一の属性値の組み合わせを満たすために，一部のセルが削除または一般化される．したがってデータセットが元々持っていた情報量が減少し，その結果，匿名化データを用いて構築した機械学習モデルの予測精度は，匿名化前のそれを用いて構築したモデルよりも下がってしまう傾向にある．よって，予測精度向上を目的とした研究が数多くある [2][3][4][5].

Rodriguez ら[6]は，機械学習の入力データに k-匿名化が適用された際の分類精度の劣化度合いを Adult など主なオープンデータセットに対して検証し，精度劣化への匿名化の影響は少ないとした．この報告ではデータセットの属性の一部のみを準識別子として扱って k-匿名化の対象とし，その他の属性は元データのままで扱っている．しかし日本においては個人情報保護法の観点から，匿名化データ（非個人情報化を意図した加工データ）からの個人情報の復元（再識別）を防ぐため，全ての属性を準識別子として扱う．これに照らすと，一部の属性のみを匿名化対象として作成されたデータは，個人情報のままであるとみなされる恐れがある．我々は本論文で，全属性を準識別子とした匿名化データを作成して精度測定することで k-匿名化は必ずしも精度劣化が小さくないことを示した．

Inan ら[2]は，匿名化による有用性の低下を抑えるため，k-匿名化した個票データに統計情報を付加することで有用性を向上させる手法を提案した．統計情報は，k-匿名化により同一グループとしてまとめられたレコード集合毎に計算され，個票データに付加される．たとえば，表 1 は属性 A_1,A_2 を持つレコード集合 R= {r_i} (i ={1,2,⋯}) と，この個票データを k=3 で k-匿名化したデータ R’を表している．データ R’の属性 A_1 と A_2 の各レコードの値は図 1 に示す一般化木を参照してデータ R における値を変換した値になっている．Inan らはこのデータ R’に表 2 に示すような付加情報 S_1, S_2 を追加することで，k-匿名性を維持しつつ匿名化データ R’の有用性を高める手法を提案した．S_1 は A_1 の付加情報であり，A_1 の各レコードの属性値についての確率を表す．たとえば P(Masters) = 1 とは，3-匿名化によりグルーピングされたレコード集合 r’’$_1$, r''$_2$, r''$_3$ の属性 A_1 の値が，匿名化前のデータ R において”Masters”である確率を表す．同様に P(11th) = 0.66 とは，レコード集合 r''$_4$, r''$_5$, r''$_6$ の属性 A_1 の値が，匿名化前のデータ R において値”11th”である確率を表す．S_2 は A_2 の付加情報であり，匿名化によりグルーピングされたレコード集合毎に平均値と分散を付加する．μ は平均値，σ^2 は分散を表す．

表 1　データセット R とその 3-匿名化データ R’

Table 1　Data set R and its 3-anonymous generalization R’

R	A_1	A_2
r_1	Masters	35
r_2	Masters	36
r_3	Masters	36
r_4	11th	28
r_5	11th	22
r_6	12th	33

R'	A_1	A_2
r'_1	Masters	[35-37)
r'_2	Masters	[35-37)
r'_3	Masters	[35-37)
r'_4	Senior Sec.	[1-35)
r'_5	Senior Sec.	[1-35)
r'_6	Senior Sec.	[1-35)

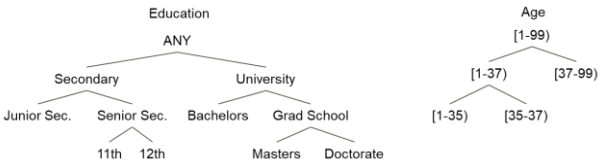

図 1　Adult データセットの k-匿名化に用いる一般化木の例

Fig. 1.　Value generalization hierarchies for Education and Age attributes of Adult dataset

表 2　レコードにグループ毎の統計情報を付加した匿名化データの例

Table 2　Sample anonymized dataset with statistics of QIs

R''	A_1	A_2	S_1	S_2
r''_1	Masters	[35-37)	P(Masters) = 1	$\mu = 35.6$
				$\sigma = 0.22$
r''_2	Masters	[35-37)	P(Masters) = 1	$\mu = 35.6$
				$\sigma = 0.22$
r''_3	Masters	[35-37)	P(Masters) = 1	$\mu = 35.6$
				$\sigma = 0.22$
r''_4	Senior Sec.	[1-35)	P(11th) = 0.66	$\mu = 27.6$
			P(12th) = 0.33	$\sigma = 20.22$
r''_5	Senior Sec.	[1-35)	P(11th) = 0.66	$\mu = 27.6$
			P(12th) = 0.33	$\sigma = 20.22$
r''_6	Senior Sec.	[1-35)	P(11th) = 0.66	$\mu = 27.6$
			P(12th) = 0.33	$\sigma = 20.22$

3. 課題

Inan らの方式により生成された k-匿名化データは，その付加された統計情報を利用して元レコードを特定できる場合があるため，個人情報漏洩の恐れがあることから日本では第三者への提供ができない可能性がある．2016 年 11 月に公示された，「「個人情報の保護に関する法律についてのガイドライン（通則編、外国にある第三者への提供編、第三者提供時の確認・記録義務編及び匿名加工情報編）（案）」に関する意見募集の結果について」の，【別紙２－４】意見募集結果（匿名加工情報編）の p.18 No.1034 の質問に対する個人情報保護委員会の回答によると，「提供元において当該 ID 等により個人情報と容易に照合できる場合には個人情報と位置付けられます」とあり，匿名化したデータであっても，そのデータの提供元事業者において匿名化前の個人情報と容易に照合できる（提供元基準で容易照合性がある）のであれば，そのデータは個人情報であり本人の同意なしには第三者提供はできない．

しかしながら Inan らの方式で生成した k-匿名化データは，そのデータ内のある匿名化グループに対応するレコードを，匿名化前のデータ内のレコードと対応付けて特定できる場合がある．Inan 方式による k-匿名化データの提供元事業者は，匿名化データの統計情報 S_i を参照して属性 A_i の属性値が x_i である確率 $P(x_i)$が 1 /（匿名化グループ数 k’（≧k））を満たす属性値 x_j を含む匿名化グループを選び，匿名化「前」のデータから，当該の匿名化グループの統計情報 S_i に含まれる属性値 x_i のいずれか（表 2 でいうと確率 P の変数である属性値 11th，12th のいずれか）を持ち，かつ S_i に含まれる数値属性 A_i の平均値 μ，標準偏差 σ の範囲に入る属性値を持つレコードを全てピックアップする．次に，ピックアップしたレコード集合からレコードを当該の匿名化グループの統計情報 Si を満たすレコードの組合せを全て作ってみる．その結果，当該の Si を満たす組合せをただ 1 つしか作れなければ，上記の属性値 x_j を持つレコードを，匿名化前のデータにおいて特定できる．

表 2 で具体的に説明する．匿名化データの提供元は $P(x_i)$ = 1 /（匿名化グループ数 k’（≧k））を満たす属性値“12th”を含む匿名化グループ $\{r''_4, r''_5, r''_6\}$に対応するレコードを匿名化前のデータ（表 1 の左側の表）から特定したいとする．匿名化データの統計情報 S_i が含む属性値のうち当該グループが含む属性値（11th，12th，μ，σ）を用いて，候補となるレコードをピックアップする．ピックアップしたレコード集合から，当該の匿名化グループの統計情報を満たすレコードの組合せが $\{r_4, r_5, r_6\}$のレコードのみであったとき，“12th”を持つという属性値は r_6 のレコードのみが保持することになるので，r''_4, r''_5, r''_6 のグループに対応するレコード r_6 を特定できる．したがって Inan 方式で統計情報を付加する際の 12th などといった変数は，匿名化前のデータのレコードと容易に照合ができ，特定の個人を識別することができるという問題がある（提供元基準で容易照合性がある）．

そこで我々は，統計情報をグループ単位ではなくデータセット単位について算出し，匿名データに付与することで提供元基準での容易照合性のない手法を提案する．

4. 提案方式

上記の課題を解決する方式として，グループごとの統計情報を付加するのではなく，個票データ全体に関する統計情報を付加する方式を提案する．付加する統計情報をデータセット単位の情報とすることにより，匿名化グループそれぞれにおける各属性の属性値の分布は，匿名化前のデータにおける対応するグループの属性値の分布と一致しなくなるので，照合したい匿名化グループに対応するレコードの組合せを匿名化前のデータから作ることが難しくなる．したがって容易には照合できず，提供元基準での容易照合性はなくなる．付加する統計情報は，具体的には各属性における属性値の「出現率」とする．この出現率の付加は，個票データの One-Hot Encoding を通して行う．そこで，まず準備として One-Hot Encoding について説明する．

4.1 準備：One-Hot Encoding

One-Hot Encoding は，機械学習分類器に入力する個票データにおける属性のうち，カテゴリ型属性のデータの前処理に用いられるポピュラーな方法であり，決定木系や線形回帰，ニューラルネット等様々な機械学習分類器に対して汎用的に入力データとして利用できるようにするためのエンコーディング方法である One-Hot Encoding の処理内容としては，個票データのカテゴリ型属性 Ai を，その属性がとりうる全ての値を属性名とする新たな属性 A_{i_j} に展開し，属性 Ai の値が j であるレコードには $A_{i\_j} = 1$ を値として割り当て，j ではないレコードには 0 を値として割り当てるというものである．図 2 に One-Hot Encoding の具体例を示す．カテゴリ型属性 A_1 の属性名を”Workclass”とし，A_1 の属性値としてレコード r_1, r_2, r_3 はそれぞれ”Private”, “State-gov”,

“Federal-gov”という値を持つ．この属性 A_1 に対して One-Hot Encoding を行うと，新たな属性 $A_{1\_1}, A_{1\_2}, A_{1\_3}$ として ”Workclass_Private”, “Workclass_State-gov”, “Workclass_Federal-gov”が作成され，対応するレコードに 1 を割り当て，対応しないレコードには 0 を割り当てる．

ID	年齢	職業	年収
1	20代	個人事業主	5万ドル以下
2	30代	フリーランス	5万ドル以下
3	40代	国家公務員	5万ドル以上
4	50代	地方公務員	5万ドル以上

One Hot Encoding（OHE）

機械学習で予測対象とする属性については One Hot Encoding 不要

ID	20代	30代	40代	50代	個人事業主	フリーランス	国家公務員	地方公務員	年収
1	1	0	0	0	1	0	0	0	5万ドル以下
2	0	1	0	0	0	1	0	0	5万ドル以下
3	0	0	1	0	0	0	1	0	5万ドル以上
4	0	0	0	1	0	0	0	1	5万ドル以上

図 2　One-Hot Encoding

Fig. 2 One-Hot Encoding

4.2 提案方式

匿名化データに One-Hot Encoding を適用し，展開されたデータを，匿名化の際に用いた各属性の一般化木（図 3）と，元データにおける各属性の値の分布情報（図 4）を用いてさらに展開する．手順は次の通り．

1. 一般化属性値（’20-30 代’，’40-50 代’，’自営’，’公務員’）の列の値が 1 であるレコードを抽出する．
2. それぞれの一般化属性値について，一般化木においてその値の配下にある葉の値（20 代，30 代，個人事業主，国家公務員など）を抽出する．
3. 抽出した葉の値の列のそれぞれに対して，「出現率」を割り当てる（図 5）．
   出現率は次の式で計算される：
   出現率 = 葉の値をもつレコード数 / 配下にある全ての葉の値をもつレコード数の和

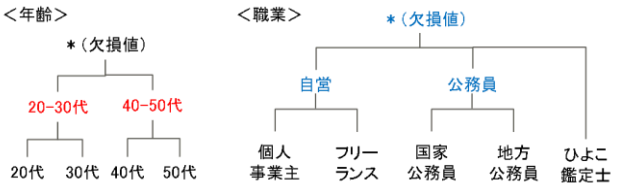

図 3　一般化木の例

Figure 3 The example of value generalization hierarchies for Age and Occupation

年齢	20代	30代	40代	50代
	10	30	31	9

職業	個人	フリー	国家	地方	ひよこ
	20	20	20	19	1

図 4　各属性の値の分布情報の例

Figure 4 The example of frequency distribution for values

匿名化データ

ID	年齢	職業	年収
91	20-30代	自営	<50k$
92	30代	公務員	≧50k$
93	40代	公務員	≧50k$
94	40-50代	自営	<50k$
201	40代	*（欠損値）	<50k$

One Hot Encoding

一般化木

属性値の分布情報

ID	20代	30代	40代	50代	個人事業主	フリーランス	国家公務員	地方公務員	ひよこ鑑定士	年収
91	0.250	0.750	0	0	0.5	0.5	0	0	0	<50k$
92	0	1	0	0	0	0	0.51	0.49	0	≧50k$
93	0	0	1	0	0	0	0.51	0.49	0	≧50K$
94	0	0	0.75	0.25	0.5	0.5	0	0	0	<50k$
201	0	0	1	0	0.25	0.25	0.25	0.24	0.01	<50k$

図 5　提案方式

Figure 5 Proposed Method

上記の手順でレコードに割り当てられた葉の値の列それぞれの出現率，すなわち一般化前の属性値それぞれの出現率は，特定の匿名化グループにおける属性値の分布ではないので，そのレコードが属する匿名化グループに対応する匿名化前のグループの照合が難しくなる．よって本方式は提供元基準での容易照合性はなくなる．

5. 評価

匿名化データを提案方式で展開し，機械学習の入力として利用したときの分類精度を評価する．比較対象として，Inan ら[2]も比較対象として用いた従来方式を用いる（従来方式 2 とする）．従来方式 2 は，提案方式の手順において「出現率」を割り当てる代わりに数値 1 を割り当てる処理に相当する．また，提案方式は欠損値をもつレコードに対しても手順に従った「出現率」を割り当てるが，従来方式 2 では何も割り当てない．

用いたデータセットは UCI データセット群のうち多くの論文で使用されている Adult データセットを用いた．Adult は年収が 5 万ドル以上かそうでないかの 2 値を予測するデータセットである．今回は，Adult データセットにおいて年収の予測に使ってよい 14 属性のうち 5 属性（age, workclass, education, capital-gain, native-counrty）を評価に用いた．これら 5 属性は Inan らが準識別子として用いた属性の一部である．

k-匿名化は多くのアルゴリズムがあるが，一般化木を用いた匿名化が可能で処理速度や情報劣化が少ないことで有名な NCP Top Down を用いた[7].

5.1 一般化木の作成方法

一般化木はいろいろな作り方がある．属性値間の意味関係をもとに作る方法，属性値の頻度に基づく方法，一般化の際の情報量の損失に基づく方法などである．本論文では次の方針で一般化木を作成した：

1. 各属性の各レコードがとる値について，Adult の目的変数（年収 5 万ドル以上かそうでないか）のオッズ（5 万ドル以上のレコード数 / 5 万ドル未満のレコード数，および分母と分子を入れ替えた値）が近いもの同士でグルーピングしていくことで，一般化によるオッズ低下を少なくする
2. 原田らの Huffman 符号木を用いた手法[8]に基づいて，頻度が近い属性値同士でグルーピングしていくことで，提案方式において頻度が少ない属性値に割り当てる出現率が，頻度の多い属性値の出現率よりも著しく小さくならないようにする．これにより頻度の少ない属性値に割り当てた出現率が，欠損値に近い扱い（つまり One Hot Encoding で展開した匿名化データにおける値が 0 に近い）となることを回避する．
3. Adult の 5 属性のうち age，capital-gain は数値属性であるが残りの属性と同じくカテゴリ属性として扱い，一般化木に組み込んだ．
4. 一般化木の階層の深さはほとんどの属性で 2 とする．Adult の age 属性のみ最大 3 とした．理由はあまり深くしすぎると実質的に元データと変わらなくなるため．

5.2 実験結果

上記の一般化木に基づいて Adult データセットを k-匿名化（k は 1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 の 11 段階．k = 1 は元データを表す）し，提案方式，従来方式 2 を用いて機械学習用の入力データに変換した時の分類精度を比較した．機械学習分類器としては 6 種類用いた：

- ロジスティック回帰（図では ‘lr’ と表記）
- 線形サポートベクター分類器（図では ‘lsvc’ と表記）
- 決定木系： ランダムフォレスト，勾配ブースティング，XGBoost，LightGBM（図ではそれぞれ順に ‘rf’, ‘gbc’, ‘xgb’, ‘lgbm’ と表記）

すべてハイパーパラメータのチューニングは行わずデフォルトパラメータで分類精度（Accuracy）を計算した．5 分割の交差検証を行い，平均値で比較した．実装は Python ライブラリを活用した．

図 6 (a)～(f)に結果を示す．すべての分類器で提案方式（青の線）が従来方式 2（橙の線）よりも精度が同じか向上した．

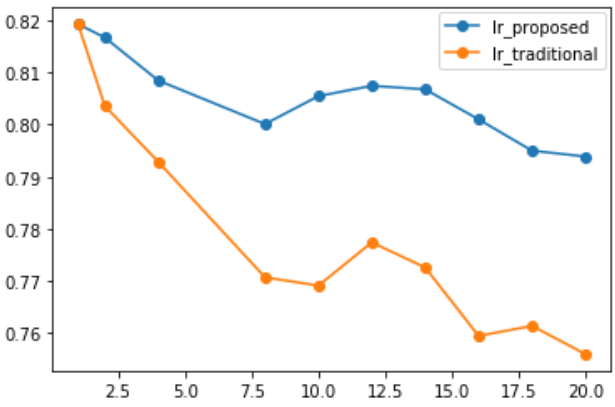

図 6 (a) 分類精度（ロジスティック回帰）

Figure 6 (a) Classification Accuracy (Logistic Regression)

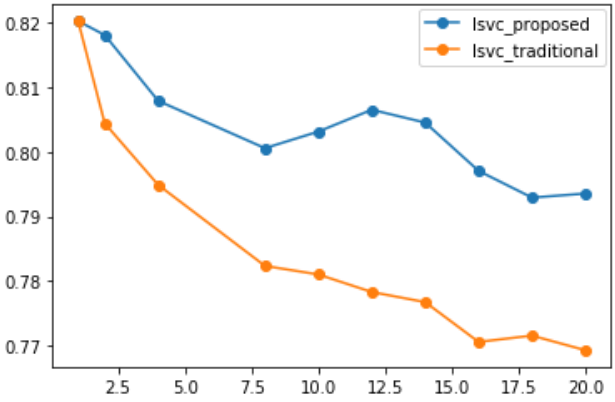

図 6 (b) 分類精度（線形サポートベクター分類器）

Figure 6 (b) Classification Accuracy (Linear SVM)

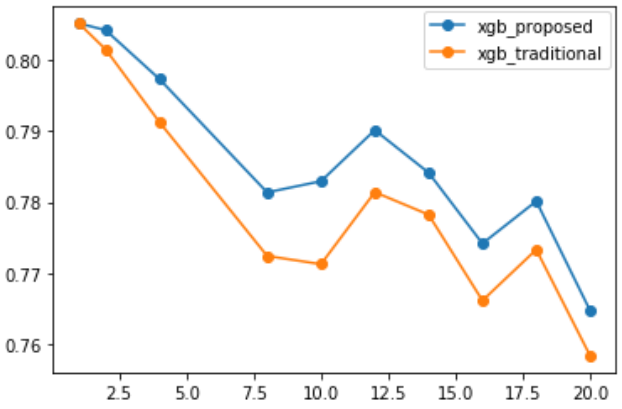

図 6 (c) 分類精度（XGBoost）

Figure 6 (c) Classification Accuracy (XGBoost)

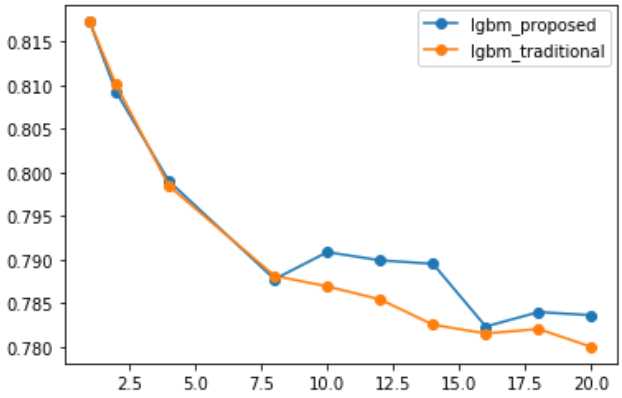

図 6 (d) 分類精度（LightGBM）

Figure 6 (d) Classification Accuracy (LightGBM)

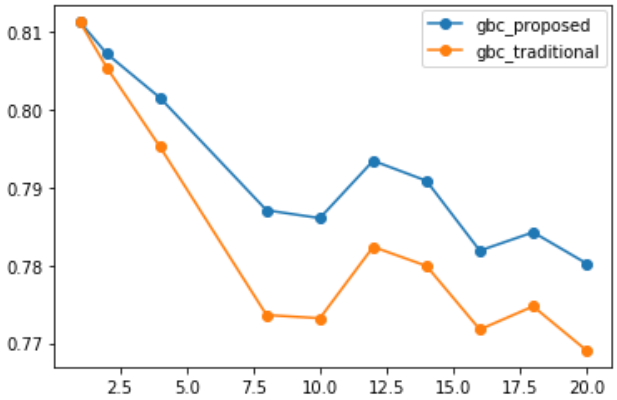

図 6 (e) 分類精度（GBC）

Figure 6 (e) Classification Accuracy (GBC)

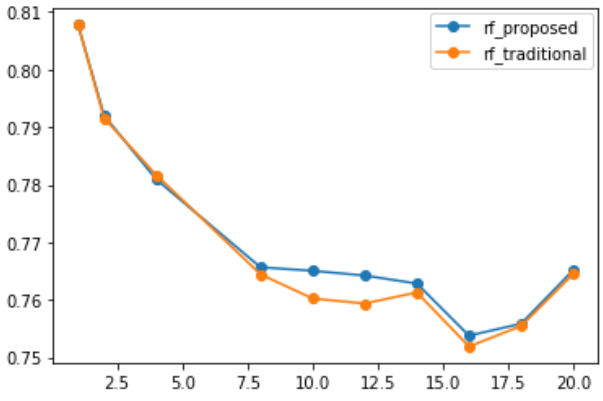

図 6 (f) 分類精度（ランダムフォレスト）

Figure 6 (f) Classification Accuracy (Random Forest)

6. 考察

6.1 分類精度の向上についての仮説

上記の実験の結果から，Adult データセットを用いて提案方式の精度向上効果について，従来方式からの向上を確認できた．この機序について仮説を述べる．今回，age, capital-gain といった数値属性の値についてもカテゴリ値として扱い，one-hot encoding に準ずる符号化して機械学習の入力データとして用いた．この場合，属性値の値集合が非常に大きくなる．属性値の値集合が大きい場合，生データ（匿名化前データ）を one-hot encoding すると，データは「0,1」を成分とする疎な行列に変換される．一方で，セル（各レコードの各属性の値）の置き換えや削除が行われた匿名化データに対して One-Hot encoding を行うと（従来方式 2），「0,1」を成分とする密な行列となる．この行列の性質の大きな異なりのため，学習されたモデルが異なり，結果として精度劣化する．

今回の提案方式によるエンコーディングは，行列が密であるという状況は変わらないが，

・行列の値が「0 から 1 の実数値」に置き換わり，

・グループに対応する箇所のみに実数値を当てはめるので，余計な「1」が生成されず，

・さらに図 4 におけるひよこ鑑定士のような属性値をほぼ 0 として取り扱うことができる

という特徴をもつので，従来方式 2 よりは密にならず，生データの状況により近くなる．この結果，若干の精度向上が得られたと考える．

6.2 ノイズ付与による匿名性の強化

提案方式は，匿名化データへの統計情報の付与を，データセット全体の単位で行うことで，グループ単位で統計情報を付与する従来方式よりも容易照合性を困難にし，提供元基準での容易照合性のない方式となっている．提案方式を適用した匿名化データを第三者提供する際，さらに容易照合性を困難にする方法として，付与する統計情報へのノイズ付与が考えられる．データセット単位での統計情報付与であっても，ある属性の属性値の出現頻度が他の属性値よりも相対的に低い場合，その属性値についての統計情報が高い，つまり提案方式を適用した匿名化データにおいてその属性値の出現率が高いレコードに着目し，そのレコードの匿名化前のレコードを探す攻撃は，出現頻度の高い属性値をもつレコードの匿名化前のレコードを探す攻撃よりも候補となるレコードが少ないので比較的容易である．そこでそのような攻撃への対処として統計情報（出現率）へのノイズ付与を行う．具体的な処理としては統計情報の元となる各属性値の出現頻度に対してノイズを付与し，その出現頻度を用いて出現率を計算して匿名化データに付与する．付与できるノイズの量に制限はないが，ノイズの量が多いほど分類精度に影響することに留意する必要がある．

6.3 統計情報を調達元サーバ経由で入手するケース

提案方式は，他組織から調達した匿名化データに One-Hot Encodiing を適用した後，属性値の分布情報を元に出現率を計算して匿名化データに付与するが，属性値の分布情報が

個人情報保護の観点から調達元のサーバ経由で提供され，差分プライバシー[9]に基づくノイズが付加される場合でも，提案方式はグループ単位の統計情報を匿名化データに付加する従来方式よりも，ノイズによる歪みの少ない分布情報をサーバから入手することができる．なぜならば調達元が，サーバが提供する統計情報をできるだけ調達先にとって意味のあるものにしたいと考える場合，差分プライバシーによるノイズの計算式における定数σはσ = Δf / n / ε で与えられるためである．ここでΔf はサーバに問い合わせられた統計情報の最大値域，n はレコード数，ε は差分プライバシーの強度（小さいほど強い）である．このとき提案方式が必要とする分布情報はデータセット全体の分布情報であるため，n が大きいため分布情報に付加されるノイズは小さいものになる．一方，グループ単位の統計情報が問い合わせられた場合，n は小さいのでノイズは強くかけられた統計情報をサーバは返すことになる．したがって，統計情報が調達元のサーバから入手するケースでは，提案方式はノイズによる歪みの少ない統計情報を匿名化データに付加することができ，機械学習の入力データとして匿名化データの有用性を高めることができる．

7. まとめ

AI を活用して高精度な分析を行うには多くのデータが必要であり，自組織のデータだけで充分な精度が得られない場合はオープンデータの活用や他組織からのデータ調達が考えられる．しかし多くのオープンデータや他組織から調達できるデータは個人情報漏洩の防止の観点から匿名化がなされている．しかしながら匿名化はデータの情報量を減らしてしまうため AI での分析精度に影響する．

本論文では，匿名化によるデータ劣化を抑制するため，匿名化前の統計情報を匿名化後のデータに付与する方式を提案し，その分類精度を評価した．結果，ロジスティック回帰分析，線形サポートベクター分類器と相性がよい傾向がわかった．

今後の課題としては，より多くのデータセットや分類器側のハイパーパラメータ調整などを通じて，よりロバストに提案方式の分類精度の向上効果や分類器との相性に関する評価を行うことのほか，ノイズ付与と分類精度劣化の関係の評価がある．

参考文献

[1] Sweeney, L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), pp.557-570, 2002.

[2] A. Inan, M. Kantarcioglu, E. Bertino. Using anonymized data for classification. Proc. IEEE Int. Conf. Data Eng. (ICDE), pp. 429-440, Mar./pr. 2009., 2009.

[3] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Workload-aware anonymization. in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), Philadelphia, PA, USA, Aug. 2006, pp. 277–286, 2006.

[4] S. Kisilevich, L. Rokach, Y. Elovici, and B. Shapira. Efficient multidimensional suppression for k-anonymity. IEEE Trans. Knowl. Data Eng., vol. 22, no. 3, pp. 334–347, 2010.

[5] Y. Jafer, S. Matwin, and M. Sokolova. Task oriented privacy preserving data publishing using feature selection. in Proc. Can. Conf. Artif. Intell., Montréal, BC, Canada, May 2014, pp. 143–154, 2014.

[6] A. Rodriguez-H., J. ESTRADA-J., D. Rebollo-M., J. Parra-A. and J. Forne. Does k-Anonymous Microaggregatoin Affect Machine-Learned Mactotrends? IEEE Access, Volume: 6, May 2018, pp. 28258-28277, 2018.

[7] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W. 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, 2006.

[8] 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, Vol.2010-CSEC-50 No.47, 2010.

[9] Dwork, C., Naor, M., Reingold, O., Rothblum, G.N. and Vadhan. On the Complexity of Differentially Private Data Release: Efficient Algorithms and Hardness Results. Proc. 41st Annual ACM Symp. Theory of Computing (STOC ’09), pp.381–390, 2009.