L-006

データ解析におけるプライバシー保護のためのデータ拡張

Data Augmentation for Privacy Preserving Data Analysis

趙　文峰†　　成　凱††

Wenfeng Zhao†　Kai Cheng††

1　はじめに

近年 IT が社会やビジネスの情報基盤となりつつ、個人レベルでもインターネットでのオンラインショッピングやモバイル端末を介したソーシャルメディアへ情報発信が行われている。人々の行動の詳細が膨大なデジタルデータとして記録され、そのデータにさまざまな分析を加えることで高度な意思決定への応用に期待が高まっている。しかし現状では組織間の壁で流通が阻まれ莫大なビッグデータはそのポテンシャルを活かしきれていない。主な原因の一つは情報の機密性の問題であり、医療データ、商品の購買履歴など個人に関する情報を含んだデータセットを不用意に第三者に公開するとプライバシーの侵害につながる危険性がある。

データに関するプライバシーを保護するため、その中で匿名化技術の研究が盛んになってきた。しかしながら、従来匿名化技法より、データの安全性が低いという問題点が実際に存在している、また、匿名化されたデータの情報量が失ったり、データの有用性を低下したりするという問題点もある。

そこで、本研究では匿名技術によりデータの安全性を保つとともに、データ拡張に基づいて、データ解析精度を高め、データの安全性と有用性を両立するというのは研究目的とする。データ拡張で作られたデータは偽データとして攻撃される心配がなく、たとえ攻撃されても、合成したデータには個人情報を含まれていないため、プライバシーなどの侵害が一切なくて、共有も容易になる。データ拡張では実データが十分ではないとき、データを合成して足していくということである。

2　関連研究

従来匿名化技法を強化する技術が存在している。南氏[1]はk-匿名性、l-多様性、t-近似性、差分プライバシーなどの代表的な PPDP のプライバシー指標およびその実現手法を解説し、通常の k-匿名化の手法の適用が困難である、トランザクションデータ、位置情報などの多次元データに対する PPDP 実現への課題、提案手法をまとめた。柿澤ら[2]は既存の Pk-匿名化、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付与することで実現されている。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する。また, ベイズ推定を用いた再構築法を適用し、攪乱されプライバシーが保護されたPk-匿名化データから統計的に有意なクロス集計結果を得ることができるかを検証する。さらに、属性値が予めグループに分類されたデータにおいても同様に再構築法を適用し、クロス集計結果を比較することで提案手法の優位性を示した。

データの解析精度を高めるため、データ拡張より、解析精度を向上する。河野ら[3]深層学習による画像分類、大量の学習サンプルが必要となるが、ラベル付けのコストは高く、さらに著作権やプライバシーの問題で十分な学習サンプルを収集できないこともある。少数の学習サンプルから Generative Adversarial Networks(GAN)を用いてデータ拡張する方法を提案した。また、GAN でデータ拡張するではなく、収集した少数サンプルで学習した畳み込みニューラルネットワークを用いて、生成サンプルの偽陽性や真陽性を事前に判定し、学習に用いるサンプルを選別する方法導入している。実験で少数サンプルの状況を模擬するために、データセットの 10%だけをデータ拡張に用い、GAN を改良した深層畳み込み GAN および Stacked GAN に対して評価した。また、比較評価には、Stacked GAN を用いて生成したサンプルのうち真陽性と判定されたもの学習に用いる方法が最も精度が高く、従来の幾何変形を用いたデータ拡張に対して正解率が 8.9%上昇することが確認できた。

3　提案手法

本研究では元データが従来匿名化技法だけより処理すると、この場合は安全性が低いと考えとして、データ拡張により、元データから合成したデータを匿名化されたデータに加え、データの曖昧さを増やすことで、安全性を向上していると考える。また、合成データで機械学習したモデルの精度を向上しているのか確かめる。提案手法の流れは図 1 のように示す。元データ D を従来匿名化技法処理して、D１を得る。データ拡張の手法より、データD２を生成し、D１＋D２を統合することで、安全性とデータの解析精度を向上する。

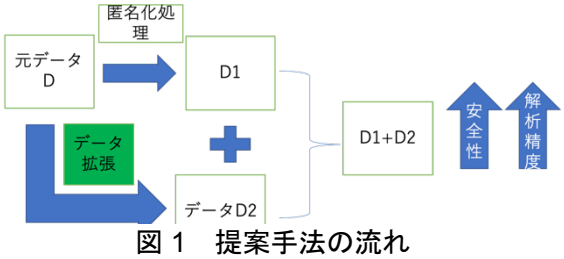


図 1　提案手法の流れ

データ拡張(Data Augmentation, DA)とは新しいデータを明示的に収集することなく、トレーニング例の多様性を高めるための技術である。そのほとんどは、既存のデータのわずかに変更されたコピーを追加するか、合成データを作成して、拡張データが正則化として機能し、ML モデルをトレーニングする際の過剰適合を減らすことを

†九州産業大学
Kyushu Sangyo University
††九州産業大学
Kyushu Sangyo University

目的としている。DA はコンピュータービジョンで一般的に使用されており、トリミング、フリッピング、カラージッターなどの手法がモデルトレーニングの標準コンポーネントである。

本研究では、データ拡張とプライバシ保護の匿名化名化技術と組合あわせることによってデータの有用性を高めることを目指す。

4　予備実験

実験で SDV による合成したデータで機械学習モデル有効性を検証する。合成したデータで機械学習モデルの解析精度を確かめるため、合成したデータで機械学習モデル有効性を検証する必要がある。

4.1 Synthetic Data Vault(SDV)

Synthetic Data Vault(SDV)[4]は、データベースの生成モデルを構築するシステムのことである。要には、データ間の関係性も考慮した上で合成データを自動的に生成してくれるシステムである。SDV で合成データを生成するメリットは、手元に実データが少量しかない場合でも、本番相当のデータをいくつでも合成できるということである。SDV によるデータセットのモデリングは以下の 4 ステップとなる。

(1) **Organize**：DB のデータをテーブルごとに別ファイルにフォーマットする。
(2) **Specify Structure**：DB のメタデータを指定する。
(3) **Learn Model**：テーブル間の関係を考慮してモデリングする。
(4) **Synthesize Data**：fit したモデルをもとに合成データを得る。

SDV により合成したデータで学習したモデルと実データで学習したモデルの精度を比較する。データの特徴量やターゲット変数の偏りを防ぐため、以下の分割交差検証を行う。

(1) データセット全体を 5 分割し、1/5 をテストデータ、4/5 を学習データとする。
(2) 2、4/5 の学習データで SDV モデルを学習させ、学習データと同数のデータを合成する。
(3) 合成データでロジスティック回帰モデルを学習させる。
(4) 学習データでロジスティック回帰モデルを学習させる。
(5) テストデータに対しての 2 モデルの精度を評価する。

上記ステップにおける 2～5 を 5 分割繰り返し、平均値を算出する。また、合成データにおけるモデル改善と実データにおけるモデル改善の対応を確かめるため、ロジスティック回帰モデルの正則化パラメータを複数設定する。

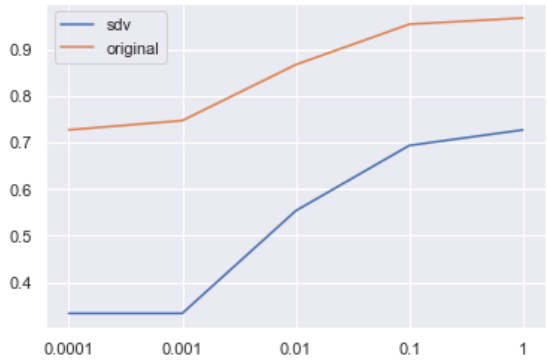
4.2 データセット

本研究では Python の scikit-learn にあるアヤメの種類と特徴量に関する iris データを使う。

4.3 実験結果

以下は、実験の結果について述べる。

図 2 パラメータのグリッドサーチ



この図 2 から、オレンジ色は元の学習データで学習したモデル、青色は SDV で合成したデータで学習したモデルである。どちらのモデルでも、正則化パラメータが増加すれば、精度の向上しているため、この正則化パラメータのグリッドにおいては精度と比例関係にある。これは、合成データ上での最適パラメータが実データにおける最適なパラメータに対応することを意味している。テストデータに対する精度は SDV で合成したデータで学習したモデルの方が低いが、ハイパーパラメータの選定においては合成データを利用しても問題がないことがわかった。

5　おわりに

本論文では，匿名されたデータの安全性と有用性に関する提案と SDV による合成したデータで機械学習モデル有効性を検証するのを述べた。今後の課題として、データ拡張したデータの有用性について評価したり、また、データの有用性が低い場合、アルゴリズムを改善したりする予定である。

参考文献

[1] 南 和宏，プライバシー保護データハバブリッシング，情報処理 Vol. 54 ,No. 9 ,Sep 2013 pp.938-946
[2] 柿澤 美穂，渡辺 知恵美，古川 諒，高橋 翼，”Pk-匿名化手法の精度改良に関する諸検討”，DEIM Forum 2015 G1-2
[3] 河野曜平，川本一彦．GAN を用いたデータ拡張[J]．研究報告コンピュータビジョンとイメージメディア(CVIM)，2017，2017(14)：1-5.
[4] Patki et al. The synthetic data vault, IEEE International Conference on Data Science and Advanced Analytics (DSAA),pp.399-410.IEEE,2016.