2ZD-04

プライバシー保護のためのデータ拡張の有用性

趙　文峰†　　成　凱†

九州産業大学†

1　はじめに

近年 IT がさまざまな分野の公共サービス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大量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しなければならない。

2　プライバシー保護技術

プライバシー保護の主要な技術として、個人を特定できる情報を削除または変更することによって、特定の個人に関連付けることができない匿名化技術である[1]。匿名化されたデータを外部と安全に共有できるため、ユーザーのプライバシーを危険にさらすことなく、他のユーザーに役立つようになる。匿名化の主な処理技術として、データマスキング、データスワッピング、データ拡張がよく使われる。

データマスキング：値が変更されたデータを非表示にする。データベースのミラーバージョンを作成し、文字のシャッフル、暗号化、単語または文字の置換などの変更手法を適用できる。たとえば、値の文字を「*」や「x」などの記号に置き換えることができる。データマスキングにより、リバースエンジニアリングや検出が不可能になる。

仮名化：個人識別子を偽の識別子または仮名に置き換えるデータ管理および匿名化方法である。たとえば、識別子「JohnSmith」を「MarkSpencer」に置き換える。仮名化により、統計の正確性とデータの整合性が維持され、データのプライバシーを保護しながら、変更されたデータをトレーニング、開発、テスト、分析に使用できるようになる。

一般化：データの一部を意図的に削除して、識別しにくくする。データは、一連の範囲または適切な境界を持つ広い領域に変更できる。目的は、データの精度を維持しながら、一部の識別子を削除することである。

データスワッピング：シャッフルおよび順列とも呼ばれる。これは、データセットの属性値を元のレコードと一致しないように再配置するために使用される手法である。たとえば、生年月日などの識別子の値を含む属性（列）を交換すると、メンバーシップタイプの値よりも匿名化に大きな影響を与える可能性がある。

データ拡張：実際のイベントとは関係のない、アルゴリズムで合成された情報である。合成データは、元のデータセットを変更したり、そのまま使用したりしてプライバシーとセキュリティを危険にさらす代わりに、人工的なデータセットを作成するために使用される。

従来技術として攻撃を防ぐため、データの情報量が失ったり、有用性を低下したりするという欠点がある。

3　提案手法

本研究ではデータ拡張を基ついて、より精度が高いデータを作るというのは本研究の目的となっている。また、アルゴリズムで作った偽データとして、攻撃される心配がない。たとえ攻撃されても、プライバシーなどの侵害も一切ない。データ拡張では実データが十分でない時、データを合成して足していくということである。本研究では主にデータ拡張手法の一つ合成データを基ついて、SDV というライブラリの上で実験を行う。

Synthetic Data Vault（SDV）[2]は、リレーショナル・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。

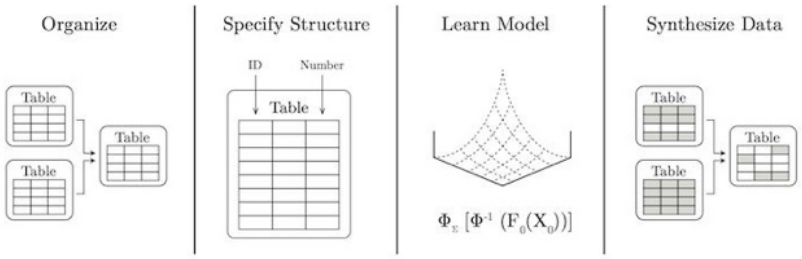

図１データセットのモデリングステップ

- **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。
- **Specify Structure**：DB のメタデータを指定する。
- **Learn Model**：テーブル間の関係を考慮してモデリングする。
- **Synthesize Data**：fit したモデルをもとに合成データを得る。

4　予備実験

実験で SDV による合成データ生成を検証する。以下は、実験内容と結果について述べる。

①データの読み込み

SDV ライブラリがデフォルトで保持している 2019 年の NASDAQ100 の株価データセットをロードし、データを確認する。

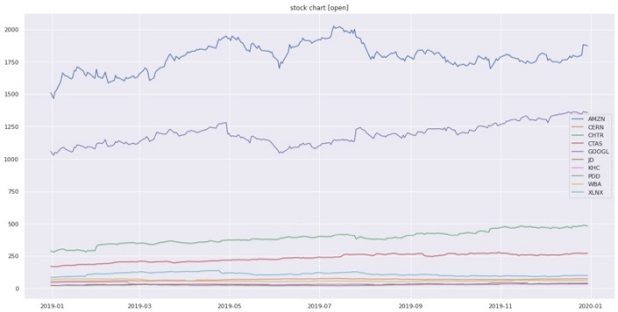

図２データセット可視化

②PAR モデルでモデリング；

SDV で時系列データをモデリングするには PAR クラスを使用する。PAR は Probabilistic AutoRegressive model:確率的自己回帰モデルの略称である。

まず、Entity カラム、Context カラム、Sequence Index を定義する。

Entity カラム：行間に依存関係が存在するグループである。このデータセットでは、銘柄ごとに行間（日付）に依存関係があるので、銘柄(Symbol)を Entity カラムとする。

Context カラム：Entity に関する属性情報を保持する変数である。このデータセットでは、時価総額(MarketCap)、業種(Sector)、業界(Industry)が Context に該当する。

Sequence Index：行間の依存関係において、順序が意味をもつような変数である。このデータセットでは、日付(Date)が Sequence Index に該当する。

次に、定義した Entity カラム、Context カラム、Sequence Index をもとに時系列データセットを PAR モデルでモデリングする。

③合成データを生成；

モデリング済みのモデルをもとに合成データを生成する。

図３original と synthetic data の比較

5　まとめと今後の課題

本論文では，プライバシーの保護の匿名化技術と SDV ライブラリの時系列データセットに対するモデリングを述べた。今後データ拡張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、アルゴリズムを改良したりする予定である。

6　謝辞

本研究を進めるにあたり，ご助言を頂いた方々に深く御礼申し上げます.

参考文献

[1] 南 和宏，プライバシー保護データパブリッシング，情報処理 Vol. 54 ,No. 9 ,Sep 2013 pp.938-946

[2] Patki et al. The synthetic data vault, IEEE International Conference on Data Science and Advanced Analytics (DSAA),pp.399-410.IEEE,2016.