2C3-3

医学・医療用擬似データに統計的性質を保って認証情報を入れるための検討

A study for medical and clinical data to put a authentication information with keeping statistical properties

城 真範 *1 Masanori Shiro　興梠 貴英 *2 Takahide Kohro　赤穂 昭太郎 *1 Shotaro Akaho

*1産業技術総合研究所 人間情報研究部門 情報数理グループ Human Informatics Research Institute, AIST　*2自治医科大学医療情報部 Jichi Medical University

In this report, we consider about methods to avoid some types of statistical modifications by appending the authentication information. We propose two ways to restore the statistical properties, which are by preprocessing the data and replacing a part of authentication information.

1. はじめに

あるデータセット X がある発信者によって発信されたことをオーソライズするために、X から発信者の（秘密にされた）キーを使う変換によって得られる認証情報 C を付与することを考える。ただし認証情報は変換 $Y = T(X, C)$ によってデータセット X そのものと不可分な形で付与し、受信者は Y のみを得る。付与にかかる計算コストは極力少なくしたい。受信者が Y を発信者に提示すると、発信者は Y と内部のキーを利用して Y の正当性を確認し、結果のみを受信者に返す。正当性確認に要する計算コストは、認証情報付与にかかるコストに比べて許容する。認証情報自体をどのようにしたらロバストに設定、付与できるかについては電子透かし技術として様々な方法が検討されている [Mohanty 1999]。しかしながら X と Y は C が付与されたぶん異なるデータであり、X の持つ統計的な性質が破壊されている可能性が高い。本稿の目的は、既に十分に安全な認証情報付与とそれを確認するのアルゴリズムがあった場合に、どのようにして X の持つ統計的な性質を保存しつつ認証情報 C を付与した Y を作るのか、その枠組みを検討することである。

この枠組みの実際的な適用例はいくつかある。例えば擬似的なデータを使ってプライバシーデータパブリッシング (PPDP) をする際、プライバシー要件を満足するように認証情報を付加した上で、公開データの情報の有用性を最大化するために統計的性質を保存したい場合である [Wang 2009, Fung 2010]。あるいは、医学・医療データを含む、ある擬似的な複雑データを生成するシステムが、十分に高度化され、実在のデータに近いデータを出力しうる場合である。この擬似的なデータは悪意ある使用者が実在のデータと偽って利用する可能性があるが、十分高度なシステムで生成された擬似的なデータは、使用者がこれを改変して利用することは、かえってデータのリアリティを損ねることになるだろう。そこで使用者がデータを改変せずに利用すると仮定するならば、そこに取り去ることの困難な認証情報を付加することで、データがシステムによって生成されたものかどうかを確認することができ、使用者が実在のデータと偽って利用することを抑制できる。

連絡先: shiro@ni.aist.go.jp

2. 提案方法

2.1 発信時の処理

データセット X が n 個のデータ $\{x_i\}(1 \leq i \leq n)$ であるとする。利用者は $\{x_i\}$ の任意の部分を抽出して利用することができる。$\{x_i\}$ のどの部分が利用されても、それがある発信者のものであることを確認できるようにするため、各データ x_i に対応した認証情報 c_i を x_i に付加する。c_i を計算するためには、十分に安全なエンコードと、エンコード結果からデータの正当性を判定する判定アルゴリズムが既にあるものとする。ただし c_i を求めるアルゴリズムは安全であるためデコード方法は存在しない。エンコードのアルゴリズムは決定論的であるが、十分に安全であるので各オリジナルデータ x_i と対応する c_i との関係はランダムに見える。すなわち $\{c_i\}$ はフォーマット構造以外は統計的に一様分布と見なせる乱数で構成されているように見える。

c_i は通常整数値であるが、正当性確認アルゴリズムが適切に実装されるなら、実数値や構造化データでも構わない。発信者は X を内部に保存しておらず、受信者からのデータ提示を受けて、自身のキーと何らかの正当性判定アルゴリズムを使ってオリジナルデータの各数値がそれぞれ正当なものであるかどうかを確認する。

ここで、c_i を単にデータに付加するだけでは、認証情報だけを削除することも容易であるから、オリジナルデータに何らかの変換 $y_i = T(x_i, c_i)$ を作用させた $\{y_i\}$ のみを受信者に提供するものとする。変換 T は一般に $\{x_i\}$ の統計的性質を破壊するから $\{y_i\}$ には $\{x_i\}$ が持っているべき統計的性質が残っていなければならない。もちろんオリジナルデータのすべての統計的性質を残すことはできないが、ここで、受信者にとってはオリジナルデータ $\{x_i\}$ の、ある統計的性質だけが重要だとすれば、$\{x_i\}$ が持つべき、その統計的性質を $\{y_i\}$ が引き継ぐように、あらかじめ発信者が $\{x_i\}$ に加工 F をしておくことで、認証情報の付加と必要な統計情報の保存を両立させることができるだろう。すなわち、発信時のデータ加工のフローは x_i に対して以下の処理を順にすることで y_i を得る。

1. $x'_i = F(x_i)$
2. x'_i から c_i を計算
3. $y_i = T(x'_i, c_i)$

ここで変換 T と F は、着目する統計値算出関数 S について、$S(y_i) = S(x_i)$ となるように設定されなくてはいけない。

2.2 確認時の処理

認証情報 c_i は十分に安全なアルゴリズムによって得られるとすれば、前述の通りある範囲の一様乱数と見なせるだろう。その期待値を $\bar{c}$ とする。変換 T と F は、c_i の違いを除けば互いに逆変換となっている。これが完全に逆変換となれば自動的に $S(y_i) = S(x_i)$ を満たすが、c_i にデコーダがないため最良推定値として $\bar{c}$ を使うこととする。すなわち変換 F に x_i 以外に引数 $\bar{c}$ を与え、$x'_i = F(x_i, \bar{c})$ とする。

例えばデータセット X が小数点以下 s 桁の実数値時系列であるとき（以下では認証情報 c_i が $[0, 9]$ の自然数であるとし、$c'_i = c_i \times 10^{-(s+1)}$ を利用して議論する。医学・医療データの場合 s は 2 から 4 程度であることが多い）簡単な $T(x_i, c_i)$ としては、

$$T(x_i, c_i) = x_i(1 + c_i)$$

が考えられる。この場合 F は、

$$F(y_i, \bar{c}) = y_i \frac{1}{1 + \bar{c}}$$

である。

さて、発信者が受信者から $\{y_i\}$ を受け取り、その正当性を確認するためには、発信時に c_i を $\bar{c}$ で代用したことによる不定性を考慮しなくてはいけない。本提案手法では、c_i が離散値であるときに、総当たりで確認を行い、一つでも正当性が確認されれば確認されたと見なす。この方法で確認の安全性を得るためには変換 T の性質について考慮する必要がある。

変換 T は一般に逆変換が定まるものなら何でも良いが、簡単な例としては単調関数である。例えば上記の F や、$x_i^{1+c'_i}$、定義域を 0 から $\pi/2$ とした正弦関数でも利用はできる。しかし一般には関数 $T(x)$ の傾きが小さなところでエンコードの安全性が崩れるため、線形関数に近い形状で複雑な挙動をもつ関数を選択するべきである。

2.3 厳密な平均値と標準偏差の保存

本提案手法では、認証情報付与にかかる計算コスト（発信時の計算コスト）は極力少なくしたいため、F には一定値である $\bar{c}$ を使う。もしも c_i の情報量が x_i の情報量に対して十分小さければ $\bar{c}/c_i$ の影響は十分小さくできると考えられる。しかしこの影響が無視できない場合には n_1 個の認証情報 c_i を破壊し、設定値 e_i に置き換える必要がある。

ここで統計量の中でも特に重要な平均値と標準偏差について、それを正確に保存することを考えるならば、問題はつまり、

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n-n_1} T(x_i, c_i) + \sum_{i=1}^{n_1} T(x_i, e_i)$$

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n-n_1} T(x_i, c_i)^2 + \sum_{i=1}^{n_1} T(x_i, e_i)^2$$

を満たす設定値 e_i を求めることに帰着する。もし認証情報 c_i が実数値で与えられるなら、e_i も実数範囲で調整することが可能で、平均値と標準偏差という 2 つの自由度を調整するためには $n_1 = 2$ で十分である。しかし確認時に総当たり検査をすることを考えれば c_i は離散値で与えられるべきである。すると、e_i は離散値の範囲で探索することになり、整数計画問題を解く必要がある。すなわち $n_1 > 2$ である必要がある。ただし、認証情報が m 個の整数値で与えられ $n_1 = 3$ であれば、m がさほど大きくない場合、m^3 の探索空間を全探索することで最適な設定値 e_i を整数範囲で探すことは、不可能なことではないだろう。

3. 実験と結果

3.1 実験

個人情報を含む医療データは特に厳密な扱いが必要で自由な利用が難しい。そこで我々は統計的性質が明らかで自由に利用できる医学・医療用擬似生体データ（擬似データ）を生成するシステムを研究・開発している [Morita 2014, Shiro 2015-1, Shiro 2015-2]。このシステムはオンラインサービスを前提としているため高速であることが必要である。ゆえに、許容されるならば整数計画問題を解かずに単純な変換 F だけでシステムを完結させたい。

今回は、将来的に当該システムに組み込むため、データ列の桁数、長さと c_i を期待値に置き換えることによる揺らぎの効果を調べた。極端な例として元データに乱数列と正弦波時系列を選択し、桁数を 2 から 5 まで、長さは 100, 1000, 10000、変換 T としては 2.2 に示した $T := x_i(1 + c_i)$ を利用した。乱数はメルセンヌツイスタを用い、C++言語にて Linux 上に実装した。

3.2 結果

以下では横軸はすべてデータ列の桁数で、縦軸は $\{y_i/x_i\}$、すなわち c_i を期待値に置き換えることによる揺らぎの効果である。10000 回の試行の平均を示した。

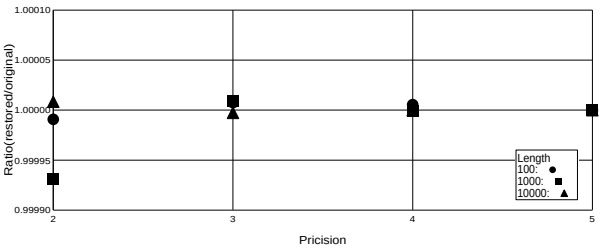

Fig1: 平均値の変化（乱数データ列）

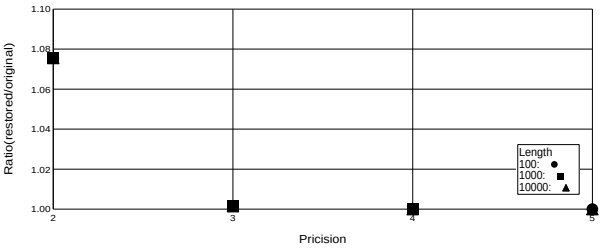

Fig2: 標準偏差の変化（乱数データ列）

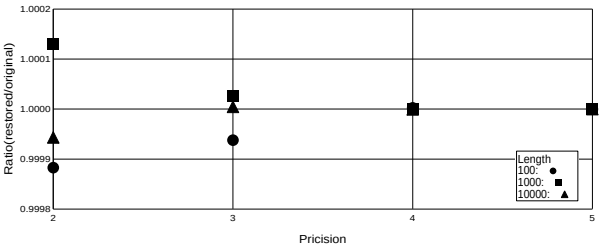

Fig3: 平均値の変化（正弦波）

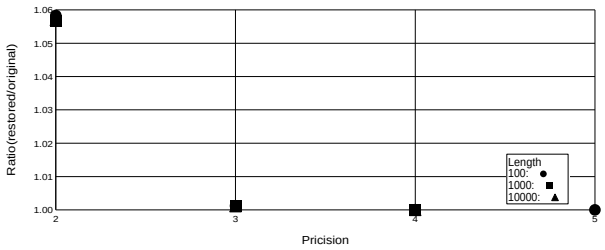

Fig4: 標準偏差の変化（正弦波）

桁数が 3 桁以上であればデータ列がどのような形でも平均値、標準偏差とも c_i を期待値に置き換えることによる揺らぎの効果は小さいことが分かる。桁数が 2 桁の場合には、発信時に単純な変換 F を行った後、特に標準偏差を調整するために数個の認証情報を適切に変更すべきである。

4. 終わりに

データに認証情報を付加することによる統計的変動を吸収するために、認証情報を得る前のデータに加工をし、かつ認証情報の一部を置き換えることで統計的変動を抑制する方法を検討した。特に認証情報を得る前のデータに加工について、その限界を調べた。将来的には、実際の認証情報付与システムを用い、また平均と標準偏差以外の統計量についても対応する方法を考えたい。

参考文献

[Mohanty 1999] Mohanty, Saraju P. "Digital watermarking: A tutorial review." URL: http://www. csee. usf. edu/ smohanty/research/Reports/WMSurvey1999Mohanty. pdf (1999).

[Wang 2009] Wang, Jian, et al. "A survey on privacy preserving data mining." Database Technology and Applications, 2009 First International Workshop on. IEEE, 2009.

[Fung 2010] Fung, Benjamin, et al. "Privacy-preserving data publishing: A survey of recent developments." ACM Computing Surveys (CSUR) 42.4 (2010): 14.

[Morita 2014] Morita, Mizuki and Shiro, Masanori: Proposal of methodology for development of pseudo clinical data generator, 医療情報学 34, pp.898-901, 2014.

[Shiro 2015-1] 城 真範, 森田 瑞樹: 医療用疑似データ生成器のカオス時系列への応用. 電子情報通信学会技術報告 NP2014-146, 2015.

[Shiro 2015-2] 城 真範, 興梠 貴英: 医学用疑似生体データの生成, 第 35 回医療情報学連合大会論文集 2015.