2M5-2

医学研究・教育用疑似データの作成

Pseudo data for medical study and education

城 真範 *1
Masanori Shiro

*1産総研 人間情報研究部門
Human infomatics RI, AIST

Useful pseudo data generator that is used in case that real data were not necessary in the medical and other fields, are discussed and partially implemented. The implementation consist of four parts included generating engine and interface layer to input difference equations with random variables. We used perl and some of free libraries for our system. Applying for actual uses, we will develop proposed filter programs for modifying time series in future.

1. はじめに

臨床医学では様々なデータが必要である。治療・研究対象とする患者の生体データのみならず、研究提案段階で使われる説明用資料、データ解析のためのプログラムを外注する際の添付資料等である。研究だけでなく、教育用資料、製薬企業等の広告資料も含めれば、医学領域が扱うデータは多岐にわたる。しかしながら、実際の人間に由来するデータを利用する場合、個人情報保護の観点から、ほとんどの場合、機関の倫理審査が必要である。倫理審査においては通常数ヶ月程度の審査期間が必要であり、また当然ながら、データを利用できる研究者の範囲や利用目的逸脱することは基本的に許されない。

だが、論文などデータの信憑性自体が問われる場合以外では、必ずしも実在の人間のデータを必要としないことも多い。例えば、上述で例示したいくつかの場合は、データは単に添付された参考資料でしかなく、倫理審査を通して正確なデータを利用することに医学的・科学的な意味は薄い。また臨床研究であっても、研究の進展等によって予期しない開発案件が発生した場合などは、解析ソフトの外注のためだけに倫理審査を待っていては貴重な時間資源が無駄になる。かといって具体的なデータを示さずに仕様を示すだけでは、複雑な解析ツールの開発において齟齬が生じやすい。研究提案の段階においても、視覚的なデータを用いた説明の幅が狭まることは、効果的なアイデアの伝達において不要な困難を加えていると言える。

こうした場合、何らかの疑似的なデータを利用することには意味がある。あらかじめ擬似的なデータであると分かっていれば、個人情報をリスクに晒す危険性はなくなり、従って倫理的な問題を原理的にクリアできるのである。

同様の事案は、医学領域のみならず、センシティブ領域と呼ばれる諸分野（金融、政治的見解、信教（宗教、思想および信条)、労働組合への加盟、人種および民族、門地および本籍地、保健医療および性生活、犯罪歴等)、あるいは特許申請中の事案を議論する際などでも発生しうる。こうした領域でデータを扱う場合、倫理的な配慮は共通の課題であり、従って擬似的なデータの利用は医学領域以外にも個人情報を扱う諸分野において有効であろう。

ところが、現状では、各固有領域に特化したモデルに基づいた擬似データが生成されるか、もしくは必要に応じて簡易的なプログラムを制作して疑似データを得ているかのどちらかが多い。それらのプログラムには共通的な部分があるはずだが、領域固有の知識と共通的な部分の峻別が難しく、その点を指摘した先行研究もあまり見あたらない。

そこで本報告では、固有のモデルや知見にとらわれることなく、分野横断的に擬似的なデータを生成するための基盤的な枠組みを提示することとした。またその一部を実装したのであわせて報告する。

連絡先: shiro@ni.aist.go.jp

2. 方法

疑似データの基本的な方法は、外部から時系列が持つべき統計量（平均や分散）を指定し、それに従ったデータ点を乱数によって繰り返し発生させることである [1]。しかし、この方法は統計的に独立した乱数列を発生させるため、連続性をもつデータ生成には適応しない。そこでデータ点同士の差分を乱数によって発生させることが行われる。この方法では連続性のみならずトレンド（上昇あるいは下降傾向）も表現可能である。本提案手法は、差分を差分方程式に拡張することで、生体データに頻出する複雑時系列（カオス、SNA 等）など様々な性質の時系列も発生可能にするものである [2]。

複雑時系列においては、乱数的要素の導入に大きく分けて二つの種類があり、力学ノイズと観測ノイズと呼ばれる。ここでは力学ノイズを差分方程式の各係数にかかるノイズであるとし、観測ノイズは観測値に対して力学系とは独立に加算されるノイズであるとする。これらを独立に設定可能とし、単純な無相関乱数も、完全な決定論的力学系も表現できるようにした。なお、ユーザがある実データから設定に必要な差分方程式を得るためには、まずデータ間の差分列を計算し、それを適切な関数で近似するだけで良い。

提案手法では、ユーザが両方のノイズ量を適切に制御することで、決定論過程と確率的過程の中間的なデータも生成できる。この手法は、疑似データの生成のみならず、既に提案されているモデルに対して人為的に乱数要素を加えることで、モデルのロバスト性を確かめたり、実データの性質を仮説検定するためのサロゲートデータを作ることもできる。

3. 実装

多様なユーザの利用端末に対して個別にコンパイルしたバイナリを提供することは困難である。そこで基本的には Web ベースのサービスとし、バックエンドは安定性の高い Linux にて運用することにした。

またユーザのニーズも多様である。例えば循環器系の医者であれば関係する検査値の推移を、できるだけ少ない入力で簡便に得たいと考えるだろう。他方、力学系の研究者であれば、自由に差分方程式を入力できることに意味がある。多くの場合は設定の簡便さと得られるデータの自由度はトレードオフの関係にあり、単一のインタフェースで様々な専門領域をカバーすることは難しい。必要に応じて細かい設定が可能で、一方でプリセットされたパラメータを使って簡単に疑似データを得ることもできるためには、様々な種類のインタフェースが提供されるべきである。そこで本実装では、インタフェース部分とデータ生成器は完全に分けて開発することにし、JSON 形式の中間ファイルを介することで階層化された複雑な設定を受け渡しできるようにした。インタフェース部分は Perl 言語にて実装した。将来的には分野に応じた様々なインタフェースを並列的に実装してゆく計画である。

また、実際には単純な時系列データの生成だけでは実用的でない。生成された時系列データを目的に応じて様々に加工する必要がある。例えば、インデックスの付加、欠損データの作成、補完、変動相関を持つデータの生成などである。これらは、作用させる順番によっても結果が異なることがあるため、生成器本体ではなく、必要に応じてユーザが作用させるフィルタとして別に実装することとした。フィルタは生成器本体とは独立しているため、疑似データのみならず、倫理上の問題がクリアできれば実データを通すことも可能である。

以上をまとめると、システムは次の 4 種類のモジュールから成る。

- インタフェース部（複数）: ユーザの入力を受け付け、中間処理プログラムに渡す。Perl で実装。
- 中間処理プログラム（単一）: データ生成器とフィルタに与えるための JSON 形式ファイルを生成し、データ生成器やフィルタを内部的に起動し、一連の処理後にデータをユーザに返す。Perl で実装。
- データ生成器（単一）: 中間処理プログラムの作った JSON 形式ファイルをロードし、与えられたノイズ付き差分方程式を解釈して指定数のデータ点列を生成する。生成データは中間処理プログラムに返す。C++にて実装。乱数の発生は標準ライブラリのメルセンヌツイスタ、JSON 解釈に `boost` ライブラリ群の中の `property_tree` 、数式解釈に GiNaC ライブラリを利用した。データ生成器における確率変数のタイプは現状で一様分布か正規分布である。確率分布はこれらの結合として式で与えることができるので、潜在的には RBF と同じ表現力をもっている。特に対数正規分布は `exp()` 関数で与えることができる。
- フィルタ（複数）: 中間処理プログラムからデータ点列とパラメータを受け取り、指定の処理を行って中間処理プログラムに返す。C++にて実装（予定、一部実装）。

上記の中でフィルタについてはまだ完全に実装できていない。

4. 展望

インタフェースについては今後規格を固定し、簡便にデータを得られるように医学的諸領域の知見をプリセットしたヴァージョンやその他の領域固有の知見に対応できるように改良する予定である。一方、得られた擬似的なデータに対して、それらをより実際的なものとするため、次節に示すいくつかのフィルタが制作されるべきである。

4.1 フィルタ

各種フィルタの実装は今後の課題である。実用上の要請から、最低でも次のフィルタは必要である。フィルタはデータ生成器とは完全に独立しているため、疑似データ以外での利用も可能である。実データに対してフィルタを通すことで、より真実みのある秘匿データを生成することも可能となるはずである。

- 欠損値生成 : 乱数を使った欠損、0 で埋めた欠損、数値部分に文字列を入れた欠損、巨大数で埋めた欠損、NA で埋めた欠損などを入れる。
- 補完：指定した n 次元関数やベジエ曲線にて点列のあいだを補完する。
- 変動相関生成：血糖値と HBA1c など、与えられた時系列に（遅延を含む）変動相関を持つデータを擬似的に作り出す。
- 位相破壊 : FFT サロゲートデータの生成器と同じ。データをフーリエ変換し、位相成分だけをランダム化して、フーリエ逆変換を行う。なお、ランダムシャッフルサロゲートに相当するデータは、時系列生成部で直接生成可能である。
- ヒストグラム化：ヒストグラムを返す。
- 特徴量抽出 : 与えられた時系列の平均、分散、微分平均、微分分散、モチーフ抽出等を行い、結果を返す。
- 構造時系列形成：複数の時系列データを使って一定の擬似的な繰り返しパターンを作る。心電図波形、脳波、体重日変動、性ホルモンの月周期変動など擬周期的なデータを生成する際に利用する。
- 離散化：離散値・記号力学系への対応。
- インデックス付加：異なる起源のデータを列方向に並べることで、時間経過に対しても乱数性を入れられる。心電図の R-R 間隔のカオス性などはこれを使って生成可能である。

4.2 フィルタ以外

次の点の改良もまた、既知の課題である。

- 再現のため、乱数の種を与えられるようにする。
- 生成とフィルタをバッチで行うための実装。
- ユーザの指定したヒストグラムに従った乱数の生成。
- 連立方程式型差分方程式への対応。
- 可視化とシンプルなインタフェース

謝辞

本研究は科研費（課題番号 : 25730154）により助成された。

参考文献

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

[2] 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, 信学技報 NLP2014-146