4B-05

確率的な再サンプリング法に基づく非線形系に対する相関係数計算

佐藤 哲 †

NHN テコラス株式会社 †

1. はじめに

データマイニングにおいて，変数の間の相関関係や独立性を確認することは重要な研究課題である．ところが，データの多様性が増大しているにも関わらず，変数間の関係が線形であることを条件としてデータ分析を実施することが旧来の手法であったため，非線形関係を検出する研究が行われている．

本報告では，変数の間に非線形を含む相関関係がある場合に，入力データに対し再サンプリング法を用いて分布を推定し，分布間の相互情報量を計算することで，相関係数を求める手法について述べる．

2. 非線形系に対する相関係数

本報告では簡単のため，離散的な値を取る 2 変数間の相関について論じる．2 変数のデータ系列 $\{(x_0, y_0), (x_1, y_1), \cdots, (x_{n-1}, y_{n-1})\}$ が与えられた場合，変数 x と y の間に関連性があるかどうかを推定する基本的な指標は次のピアソンの積率相関係数である：

$$r = \frac{\sum_{i=0}^{n-1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n-1}(x_i - \bar{x})^2 \sum_{i=0}^{n-1}(y_i - \bar{y})^2}}$$

式から見て取れるように，ピアソンの積率相関係数は入力データ (x_i, y_i) 及びその平均値 $(\bar{x}, \bar{y})$ より計算でき，その値は線形な相関関係が弱いほどゼロに近い値を取る．しかし，ピアソンの積率相関係数は計算の簡単さや直感的な理解が容易い反面，線形の関係以外は検出できないこと，入力データに外れ値が含まれている場合には精度が下がることが知られている．

非線形系の入力データに対する相関係数はピアソンの積率相関係数と同様で，値がゼロであれば 2 変数は独立しており関連性は無く，関連性があり相関が強いほど大きな値を取る．様々な係数が提案されているおり，例えば多重解像度で離散的な相互情報量を計算することにより相関係数を計算する MIC(Maximal Information Coefficient)[1], カーネル法に基づきヒルベルト空間で相互共分散を計算することにより相関係数を計算する HSIC(Hilbert Schmidt Independence Criterion)[2] など多くのものが知られている．これらの手法は，ピアソンの積率相関係数に比べると良好に非線形な変数間の相関関係を数値的に計算可能であることが示されている．しかし，入力データの量が少なかったり外れ値が多かった場合には精度が落ちることがある．

そこで本発表では，再サンプリングにより入力データの分布を推定することで，汎化性能を向上させた相関係数の計算手法を提案する．本手法はノイズを考慮しているため外れ値への耐性が高く，分布の線形性やガウス性は仮定しないため非線形系にも対応する．

前述の 2 つの手法のアプローチを比較すると，以下のようになる：

- MIC - 無秩序状態からデータ系列がどれくらい離れているかを数値化する
- HSIC - 秩序状態からデータ系列がどれくらいどれくらい離れているかを数値化する

本発表で報告する内容は，一様分布から生成されたサンプルデータと入力データがどれくらい離れているかによって相関係数を計算するアプローチであり，MIC の考え方に近い．以下のアルゴリズムによって計算される値を，便宜上 R-AMIC (Resampling-based Approximated Mutual Information Coefficient) と呼ぶ：

1. 以下の処理を入力データの数 n だけ繰り返す．
   (a) 一様分布に従うサンプルデータを m 個発生させる．

   $$\boldsymbol{p}_k^{(i)} = (\max(\boldsymbol{q}) - \min(\boldsymbol{q}))\boldsymbol{v} + \min(\boldsymbol{q})$$

   ここで，$\boldsymbol{v}$ は $(0, 1]$ の範囲の一様分布に従う乱数，$\max(\boldsymbol{q})$ 及び $\min(\boldsymbol{q})$ は入力データの座標成分の最大値と最小値を表し，$0 \leq i < m$，$0 \leq k < n$ である．
   (b) $\boldsymbol{p}_k^{(i)}$ に対し，入力データ $\boldsymbol{q}_k$ との距離及びノイズ ω に基づく確率で再サンプリングする．

   $$\boldsymbol{q}_{k+1}^{(i)} = resampling(\boldsymbol{p}_k^{(i)}, \boldsymbol{q}_k, \omega)$$

   ここで，再サンプリングは同じサンプルの複数回選択が可能な復元抽出とする．

Calculating Correlation Coefficient for Nonlinear System based on Probabilistic Resampling Method

†Tetsu R. Satoh

†NHN Techorus Corp.

2. $\{\boldsymbol{q}_0,^{(0)}, \boldsymbol{q}_0^{(1)}, \cdots, \boldsymbol{q}_0^{(m-1)}, \boldsymbol{q}_1^{(0)}, \cdots, \boldsymbol{q}_{n-1}^{(m-1)}\}$ の座標成分の相互情報量を計算する．

$$\text{R-AMIC} = MI(\boldsymbol{q}_k^{(i)})$$

相互情報量を計算する点は MIC と同じであるが，ステップ (b) において入力データの尤度を計算していることで，MIC の多重解像度における相互情報量計算を不要としていることが特徴である．

3. 相関係数値の比較実験

ここでは，人工的に作成したデータ及び調査によって得られた現実世界のデータに対し，ピアソンの積率相関係数，R-AMIC，MIC，HSIC の値を計算した結果を紹介する．作成したデータは，以下の 9 種類である：

1. $y = x + v$（直線）
2. $y = x + 50v, (50 - x) + 10v$（クロス）
3. $y = (x - 25)^2/25 + 10v$（2 次関数）
4. $(x, y) = (50v, 50v)$（一様分布）
5. $(x, y) = (20\cos(50v) + 20, 20\sin(50v) + 20$（円形）
6. $y = 20\sin(25v) + 20v + 20$（三角関数）
7. $(x, y) = (10v, 10v), (10v + 30, 10v + 30)$（擬似相関）
8. $y = 10v + 25$（y = 定数）
9. $(x, y) =$ (年齢, 港区の男性の人口)(乱数混入無しデータ)

これらの入力データに対し，各相関係数の値を計算した結果を図 1 に示す．ノイズが混入しているデータや閉じた曲線が元になっているデータ，周期を持つデータなどがある．この計算結果を，以下の 3 点について確認する：

a) 相関関係が存在しない「一様分布」「擬似相関」「y=定数」で低い値を示しているか
b) 相関関係が存在する「直線」「クロス」「二次関数」「円形」「三角関数」で高い値を示しているか
c) 相関関係が存在し，ノイズの含まれていない「人口データ」で高い値を示しているか

まず a) については，R-AMIC は全て条件を満たしており，MIC と HSIC は「擬似相関」を満たしていない．次に b) 及び c) については，相関関係が存在しない「一様分布」と比べ高い値を示しているという意味ではどの手法も条件を満たしている．ただし MIC と HSIC では，「クロス」「二次関数」「円形」に対し，相関関係が存在しない「擬似相関」よりも低い相関値が出ている．これらの結果は，R-AMIC の優位性を示唆している．

以上の実験では，ピアソンの積率相関係数及び R-AMIC の値は *Wolfram Mathematica* 上での実装を用いて計算した．また，MIC は minerva† を，HSIC は dHSIC†† を用いて計算しており，パラメータは全てデフォルト値である．また，人口データについては，港区オープンデータ ††† より，2017 年男性年齢別人口の 0 歳より 60 歳のデータを参照した．

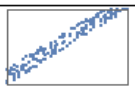
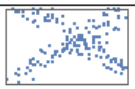
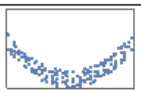
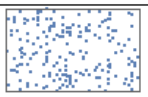
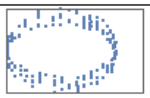
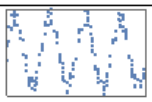
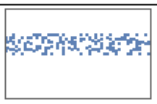
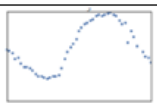
形状	プロット	ピアソン	R - AMIC	MIC	HSIC
直線		0.97	1.12	0.93	0.06
クロス		-0.03	0.67	0.37	0.02
2 次関数		0.06	0.67	0.73	0.04
一様分布		-0.05	0.55	0.17	0.01
円形		0.06	0.64	0.51	0.03
三角関数		-0.15	0.58	0.85	0.02
擬似相関		0.93	0.55	0.84	0.06
y = 定数		-0.01	0.43	0.11	0.01
人口データ		0.59	0.97	1.00	0.06

図 1　相関係数値計算例

4. おわりに

本報告では，ノイズを考慮した再サンプリング法により入力データの分布を推定し，座標成分の分布間の相互情報量を計算することで相関係数を計算する手法を提案した．今後の課題には，計算コストの高さの解決などがある．

参考文献

[1] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, Detecting Novel Associations in Large Data Sets, Science, Vol. 334, No. 6062, pp. 1518-1524, 2011.
[2] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schlkopf, and A. J. Smola, A Kernel Statistical Test of Independence, Proc. NIPS 2007, pp. 585-592, 2007.

†https://cran.r-project.org/web/packages/minerva/
††https://cran.r-project.org/web/packages/dHSIC/
†††https://www.city.minato.tokyo.jp/opendata/jinko/nenreibetsu/index2.html