圧縮アルゴリズムを用いたデータの複雑度や類似度の推定とその応用

JST ERATO/東京大学 生産技術研究所
下川 英敏[1]

テキストのユニバーサル圧縮の手法は, 様々な計算機上に実装され, 記録や通信において広く利用されており, 非常に高い有効性および有用性が示されている. これらの手法は, 基本的にはパターンマッチングなどの手法により, 情報源の確率分布を陽または陰に推定し, 情報源のエントロピーレート近くまでの圧縮を実現している. 一方, コルモゴロフ複雑度は, 圧縮の理論的な限界を表す量としてアルゴリズム的情報理論ではよく用いられており, 情報系列の本質的な複雑度を表わす指標と考えられている. コルモゴロフ複雑度は理論的に多くの好ましい性質を持っており, データの特徴量としてこの複雑度を用いることにより, 様々なデータ解析への応用が考えられる. しかしながら, 一般に個々の系列に対してコルモゴロフ複雑を求めることは困難である.

本講演では, gzip,bzip2 や CTW などの圧縮アルゴリズムを用いて, 主に神経細胞の発火時系列の複雑度を推定し, データ解析へ応用した例を紹介した. 具体的には, 脳幹の下オリーブ核での実験データ [1] を用いた電気的結合 (gap junction) の強さと発火の複雑度の定量的な評価, 発火パターンの有意性を複雑度を統計量としたサロゲート法を用いて検定する手法 [2], 一次運動野の発火時系列を幾つかの圧縮アルゴリズムを用いて圧縮し, それらの圧縮率の変化に基づく時系列の長期記憶性などの議論 [3], などの紹介をした. さらに, 複雑度に基づく類似度 [4] を用いて, 下オリーブ核の複数の神経細胞の発火パターンからクラスタリングを行い, 物理的に近接する細胞間でより発火パターンが近いことを示した. これは, 電気的結合が下オリーブ核で優位であるという解剖学的な事実と符合する.

これらの手法のメリットは, 一般に扱いが困難な複数の点過程のデータを比較的容易に扱える点にある. しかしながら, テキストのユニバーサル圧縮のアルゴリズムが必ずしも点過程データの圧縮に最適化されているわけではない. 今後の発展としては, このような複数の点過程データに適した圧縮手法を開発し, それに基づく解析手法を導き出すことが望まれる.

参考文献

[1] E. Lang, I. Sugihara and R. Llinás, Journal of Neurophysiology, 76 (1996), 255.

[2] Y. Hirata et al., Submitted to Journal of Neuroscience Methods.

[3] Gao, Kotoyiannis and Bienenstock, Proc. of ISIT (2006), 645.

[4] R. Cilibrasi and P. Vitányi, IEEE Transactions on Information Theory, 51 (2005), 1523.

[1]E-mail: simokawa@sat.t.u-tokyo.ac.jp