幾何学的制約に基づいた高相関変数集合導出手法

Derivation of High Correlation Coefficients based on Geometric Constraints

中西耕太郎 [1*]　鷲尾隆 [1]
Kotaro Nakanishi[1] and Takashi Washio[1]

[1] 大阪大学産業科学研究所高次推論方式
[1] Department of Advanced Reasoning,
The Institute of Scientific and Industrial Research, Osaka University.

Abstract: Correlation analysis among variables are frequently used in various approaches of statistics and data mining. However, its application to the data obtained from the recent ubiquitous sensing system consisting of massive sensors is often intractable, since its computational complexity is proportional to the square number of variables. On the other hand, the strong correlations among the variables are usually sparse in various data such as small world data. In this report, we propose a novel method to efficiently estimate the correlations among massive variables under this sparseness. Its experimental evaluations show excellent performance in efficiency comparing with the direct computation of all correlation coefficients.

1　はじめに

相関解析は統計的因果推論に代表されるように，データマイニングや統計数理の様々な分野で用いられる基礎的解析手法である [1, 2]．例えば，最新鋭の発電所や自動化が進む自動車工場での異常検出 [3, 4] 等では，人間の知覚に代わる異常監視センサから得られるデータを解析し，”異常 X はセンサ A, B, C にのみ多大な影響を与える．”といった，センサの値と異常との相関関係を導出して知識ベース化することにより，異常診断の自動化が可能となる．このように，相関解析は学術的基礎手法であるのみならず，様々な工学的分野で多用されている．

相関解析は，変数を列，事例を行とした二次元配列の形式で構成されるデータを対象とする．例えば，変数個数 M，事例数 N のデータは以下の形式で表される．但し，i 番目の事例における変数 j の値を x_{ij} とする．

$$\begin{array}{c} \\ \text{事例}_1 \\ \vdots \\ \vdots \\ \text{事例}_N \end{array} \begin{array}{c} \begin{array}{cccc} \text{変数}_1 & \cdots & \cdots & \text{変数}_M \end{array} \\ \begin{bmatrix} x_{11} & \cdots & \cdots & x_{1M} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{N1} & \cdots & \cdots & x_{NM} \end{bmatrix} \end{array}$$

*連絡先：大阪大学産業科学研究所
〒 567–0047 大阪府茨木市美穂ヶ丘 8-1
E-mail: nakanishi@ar.sanken.osaka-u.ac.jp

相関解析の目的は，このようなデータから強い関連を有するすべての変数ペアを導出することである．

相関解析では，解析対称データに含まれる M 個の変数から構成しうる全ての変数ペアに関して，相関係数を計算しなければならない．即ち，${}_M C_2$ 個の相関係数を計算しなければならない．また，その都度，それぞれ N 事例分存在する各変数の値にアクセスしなければならない．故に，相関解析の時間計算量は $O(M^2 N)$ となる．故に，相関解析が解析対象とし得るデータの変数個数は，実用的観点から制限される．

しかしながら，今日，計算機の処理能力の向上，大容量記憶媒体の低価格化，情報通信技術の急速な発展等により，データの大規模次元化が進んでいる．例えば，先程述べた発電所や自動車工場では，より詳細な異常監視のために数万個を超えるセンサを含む最新鋭の異常診断設備が導入され，センサ数が爆発的に増加している．また，最近の新しい技術動向として，RFID((Radio Frequency Identification) タグによる物流管理に代表されるように，ユビキタスセンシング技術の普及により，膨大な変数によって事象を詳細にデータ化できるようになって来た．このようなデータに相関解析を適用すれば，これまでの解析対象データからでは得ることが出来なかったより詳細で有意な相関関係を知識として導出できるであろう．しかしながら，先述のように従来の相関解析では変数個数が増大すると計算量が著しく増大するため，大規模次元データに対する適用が困

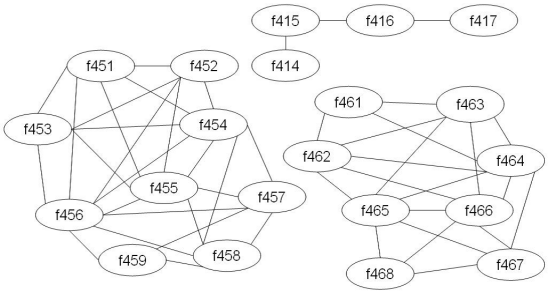

図 17: Isolet データから導出した高相関変数ペア集合

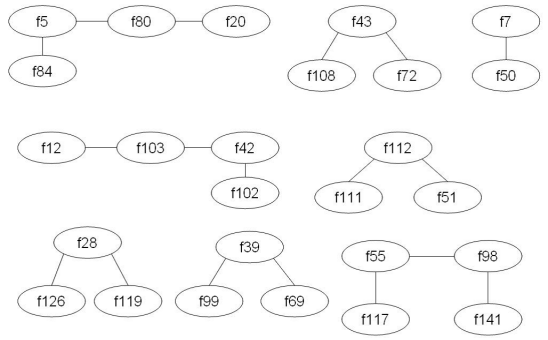

図 18: Isolet データから導出した高相関変数ペア集合

用が期待できる.

参考文献

[1] Simon, H. Causal Ordering and Identifiability. In Models of Discovery, pages 53-80. D. Reidel, Dordrecht, Holland, (1953).

[2] Blalock, H. M. Causal Inferences in Nonexperimental Research. The Univ. of North Carolina Press, Chapel Hill, North Carolina, (1961).

[3] Z. Sun, R. Miller, G. Bebis, and D. DiMeo. A real-time precrash vehicle detection system. Proc of the 2002 IEEE Workshop on Applications of Computer Vision, pp.171–176, (2002).

[4] K. Sycara and M. Lewis. From data to actionable knowledge and decision. Proc. of the 5th Int. Conf. on Information Fusion, Vol.1, pp.577–581, (2002).

[5] Newman, D.J., Hettich, S., Blake, C.L., and Merz, C.J. UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, (1998).