時系列データを対象としたクラスタリングツールの開発

1X−03　干田尾 隼人† 片山 薫‡ 鈴木 敬久‡ 太田 学‡ 石川 博‡
†東京都立大学工学部電子・情報工学科　‡東京都立大学大学院工学研究科

1 はじめに

本研究で対象とするものは、核融合をめざすための磁場閉じ込めプラズマ実験装置より得られるデータである[1]。データはプラズマを生成するための印加磁場を測定したときの電圧波形と、生成されたプラズマの密度変化を示すレーザー干渉計からの波形より成っている。現在は、数千回の実験データを人の目で分類している。この実験に対して実験データをコンピュータでクラスタリングすることで入力電圧とプラズマの密度との間の関係を短時間で調べることが期待できる。そこで、本研究では上記の実験から得られた時系列データをクラスタリングするツールを開発し、実データを用いて実験、評価を行った。このツールには複数のクラスタリングアルゴリズムが実装されているので、ユーザがアルゴリズムごとの結果を簡単に比較し、適切なものを選ぶことができる。データの形式はNetCDF[2]である。これについては2.2で詳しく述べる。また実装したアルゴリズムについては2.3で詳しく述べる。本稿ではまず開発したツールの概要を説明する。そして行った実験の概要を説明し、それを評価する。

2 クラスタリングツール<名前>

ここでは、本研究で開発したツール<名前>の概要を述べる。

2．1 概要

1つのNetCDFファイルには1つ、もしくはそれ以上の時系列データが含まれる。それらの時系列データを本ツールではユークリッド距離をとるために多次元空間上の点として扱う。例えば1つのファイルの中にサンプリング数1000の波が3つあれば3000次元の点として扱う。クラスタリングの対象となるファイルすべての点について互いの距離をとり、その距離に基いてクラスタリングを行う。距離の取り方としては、そのままユークリッド距離をとる方法と、離散フーリエ変換によって高次元データを低次元空間に落としてから距離をとる方法がある。上記2つの方法により算出されたデータ間の距離に基づいてward法[3]によりクラスタリングを行う。データをいくつのクラスタに分けるかはユーザが選ぶことができる。図1に処理の流れを示す。

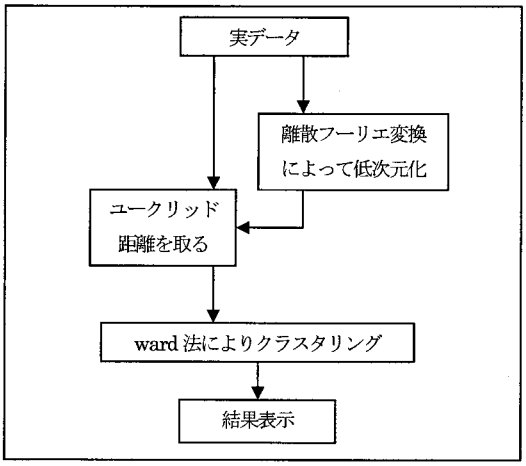

図1 処理の流れ

2．2 対象データの形式

本ツールで扱うデータ形式はNetCDF(Network Common Data Form)である。今回は1つのファイルの中から3つの時系列データを取り出し、それをクラスタリングの対象としている。3つのデータは全て8ビットサンプリングで同一スケール、同一ベースラインで同一次元(2048次元)の時系列データである。よって、1つのファイルから6144次元の1点を抽出してクラスタリングを行う。

2．3 実装したアルゴリズム

ここでは本ツールに実装されているクラスタリングアルゴリズムについて簡単に触れる。

・ユークリッド距離

次元数NのデータXとYのユークリッド距離$d(X,Y)$は(1)式で定義される。

$$d(X,Y)=\sqrt{\sum_{i=0}^{N}(x_i-y_i)^2} \quad (1)$$

Development of a clustering tool for time series data
Hayato Hidao†, Kaoru Katayama‡, Yukihisa Suzuki‡, Manabu Ohta‡, Hiroshi Ishikawa‡
†Faculty of Engineering, Tokyo Metropolitan University
‡Graduate School of Engineering, Tokyo Metropolitan University

この方法はデータによってスケールや振幅が違う場合に向かないが、本ツールの対象データはすべて同じスケールであり、振幅にも大きな変化がないことからユークリッド距離をとることに問題はないと判断した。また、短期間の振動やノイズに影響を受けやすいが、対象データには短期間の振動が少ないことから考慮に入れず、ノイズに関しても考慮しないこととした。

・離散フーリエ変換

ユークリッド距離をそのまま計算すると、データが高次元になればなるほど計算にかかる時間が大きくなってしまう。そこで、高次元データを低次元に落としこむ事が必要になる。高次元のデータを低次元に落としこむ方法として本ツールではデータを離散フーリエ変換し、その係数に基づいて距離を取ることとした。処理を軽減するために高速フーリエ変換を採用した。

・ward 法

上記の２つの方法により得られたデータ間の距離を使い、本研究では階層型クラスタリング手法の中で広く用いられるward法を使ってクラスタリングを行う。２つのクラスタ p,q をまとめてできるクラスタ t と任意のクラスタ r との距離の更新は(2)式により求まる。

$$S_{tr}=\frac{n_p+n_r}{n_t+n_r}S_{pr}+\frac{n_q+n_r}{n_t+n_r}S_{qr}+\frac{n_r}{n_t+n_r}S_{pq} \qquad (2)$$

S_{ab}　クラスタ a,b 間の距離

n_c　クラスタ c に含まれるデータ数

3　実験と評価

作成したツールを使い、そのままユークリッド距離を取った場合と低次元空間に落としてからユークリッド距離を取った場合のクラスタリング結果を比較した。実験で用いたデータは、冒頭に述べた実験での入力電圧のデータである。本実験では255個のデータを10個のクラスタに分け、そのままユークリッド距離を取った場合を正しいと仮定し、離散フーリエ変換した後に採用する係数の数を変えて適合率を検証した。適合率は(3)式で定義した。

$$\text{適合率}[\%]=\frac{\text{正しいクラスタに含まれたデータ数}}{\text{全データ数}}\times 100 \qquad (3)$$

係数は周波数の小さいものから取るものとした。今回はクラスタ数を10としたが、これは実験の結果クラスタ数が 10 を越えるとクラスタ間の距離が突然大きくなる事が分かったためである。実験の結果を図２に示す。

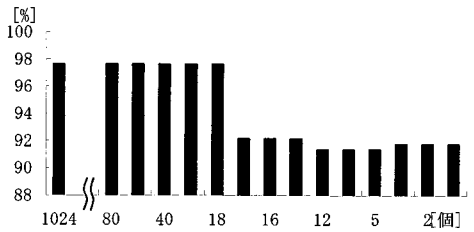

図２　採用した係数の数と適合率の関係

この結果から、離散フーリエ変換によって2048次元のデータを 20 次元程度にまで落とし込むことができる事が確認できた。処理全体にかかる時間はそのままユークリッド距離をとる場合で80～90秒、離散フーリエ変換する場合は離散フーリエ変換に要する時間を除いて 60～70秒であった。１度離散フーリエ変換しておけば、処理にかかる時間を短縮できることが確認できた。

4　おわりに

本稿ではクラスタリングツールの開発とそのツールを使って実験および評価を行った。今までは実験結果を全て人が分類していたが、クラスタリングを行うことで実験結果の分類が容易になることが期待できる。

現在は実験の入力データのみをクラスタリングしているが、出力データを含めたクラスタリングをして入出力の関係を調べることが今後の課題として挙げられる。

謝　辞

本研究を進めるにあたり、大阪大学大学院工学研究科付属超高温理工学研究施設より提供して頂いた実験データを使用した。ここに深く感謝する。

本研究の一部は、文部科学省科学研究費特定研究領域(C)(2)「情報学:A02」(課題番号:13224078)による。

参考文献

[1] H.Himura, S.Okada, S.Sugimoto, and, S.Goto; Phys. Plasmas 2, 191(1995).

[2] Unidata NetCDF http://www.unidata.ucar.edu/packages/netcdf/

[3] 菅 民郎: 多変量解析の実践(下), 現代数学社

[4] Christos Faloutsos, M.Ranganathan and Yannis Manolopoulos: *Fast Subsequence Matching in Time-Series Databases*, Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data, Minneapolis, Minnesota, May 24-27, 1994, pages 419-429.