社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－BIO－5 (11)
2006／6／15

時系列クラスタリングのためのスパイダーアルゴリズム

山村 雅幸† 亀田 祥平†

†東京工業大学大学院総合理工学研究科 〒226-8502 横浜市緑区長津田町 4259
E-mail: †my@dis.titech.ac.jp

あらまし 情報技術の急速な発展とインフラ化に伴って、さまざまなソースからの時系列データが大量に蓄積されはじめている。時系列データの適切な解析手法の整備が急務である。従来、時系列のクラスタリングにおける類似性は系列間のユークリッド距離をベースに定義されてきたが、多変数に関する情報が失われる欠点がある。本研究では、時系列クラスタリングのために、クモの生態系にヒントを得た新しいアルゴリズムを提案する。そこでは、生態系を通じたクモの棲み分けによって、時系列データの特徴点を巣の位置として抽出し、それらの特徴点の順序相関に基づいてデータ間の距離を定義しクラスタリングに役立てる。オーストラリアの手話の軌跡データを用いて、分類性能が従来の方法より高いことを実験的に調べた。

キーワード スパイダーアルゴリズム、ナチュラルコンピューティング、人工生命、時系列、クラスタリング、データマイニング

1. はじめに

情報技術の急速な発達とインフラ化に伴って、株式市況、ユビキタスセンサー、遺伝子発現プロファイルなど、さまざまなソースからの時系列データが大量に蓄積されはじめている。大量データからの知識発見をデータマイニングと言うが、時系列データの適切なマイニング手法の整備が急務である。

データマイニングのひとつの方法にクラスタリングがある。クラスタリングではさまざまに定義された類似度にしたがって、データを分類する[1]。多くのアルゴリズムが提案されている[2]。一般にクラスタリングの性能は類似度の定義の選択に依存する。

最も簡単な類似度は、時系列データを幾何学図形と見たときの各点のユークリッド距離であり、最も頻繁に用いられる。しかしながら、時系列としてみた場合、ユークリッド距離ではしばしば直感的なクラスタリングに失敗する。これはユークリッド距離が時間軸の微細な変動に非常に敏感だからである[3]。

もうひとつの良く知られた類似度は dynamic time warping である[4]。dynamic time warping はもともと音声認識の分野で 1970 年代に提案された。各データ点を多対 1 、 1 対多にアラインメントすることで、時間軸の変動にあまり依存せずに、より幾何学図形としての時系列の形を対応付ける。正確さの代償として計算量がかさむので、従来研究の多くは、dynamic time warping の計算量の低減に主眼を置いており、クラスタリングの機能面においてより優れた方法を提案することには消極的であった。さらに、従来研究の多くは、単変数時系列を扱い、多変数には注目していない。多変数では、ユークリッド距離にせよ、dynamic time warping にせよ、多変数では距離を合算してしまうために、変数間の関係が失われてしまう欠点がある。

本研究では、多変数時系列のクラスタリングのための新しいアルゴリズム「スパイダーアルゴリズム」を提案する。スパイダーアルゴリズムでは、従来手法で失われてきた変数間の関係も利用される。名前の示すとおり、このアルゴリズムは生態系におけるクモの巣の棲み分けに啓発されたものである。生命から啓発されたアルゴリズムにはニューラルネットから遺伝的アルゴリズムまでさまざまなものがあるが、いずれもダイレクトで直感的にわかりやすい応用を特徴とする。本研究でのクモの巣の使い方はひとひねりしてある。そこでは、生態系を通じたクモの棲み分けによって、時系列データの特徴点を巣の位置として抽出し、それらの特徴点の順序相関に基づいてデータ間の距離を定義しクラスタリングに役立てる。以下、アルゴリズムについて説明し、オーストラリアの手話の軌跡データを用いた計算機実験によって、従来手法との比較を行う。

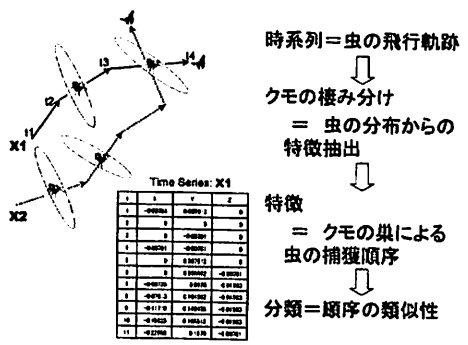


図1 スパイダーアルゴリズムによるクラスタリングの概念図

2. スパイダー・アルゴリズム

2.1. 概要

時系列のセットは次のように与えられる。

$$X = \{X_1, X_2, ..., X_N\}$$
$$X_k = (\mathbf{x}_1^k, \mathbf{x}_2^k, ..., \mathbf{x}_t^k)$$
$$\mathbf{x}_t^k = (x_{t1}^k, x_{t2}^k, ..., x_{tm}^k)$$

ここで、X_kはＮ個のうちｋ番目の時系列、$\mathbf{x}_t^k$はX_kのｔ時点のデータ点でｎ次元数値ベクトルである。

スパイダー・アルゴリズムにおける処理の流れを図２に示す。まず、前処理として時系列データの標準化を計算する。これによって、スケールの違い等による動作パラメータへの依存性を減らし、汎用化に役立てる。次に、クモの世代交代を含む、クモと虫の生態系からなる適応過程がある。ここで、虫すなわち時系列のセットが持つ特徴をクモの巣の棲み分けを通じて抽出する。適応過程は各世代で捕獲リストを出力する。捕獲リストとは虫すなわち時系列がどのクモの巣によって捕獲されたかの順序リストからなる。捕獲リストに基づいて時系列の類似度が計算される。ここまでが、スパイダーアルゴリズムである。計算された類似度に既存の任意のクラスタリングアルゴリズムを適用することで最終的なクラスタリングを得ることができる。

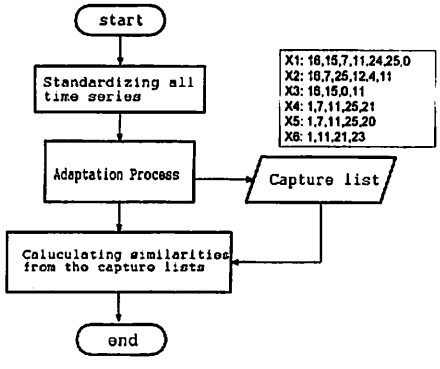


図２ 処理の流れ

2.2. 虫

既に述べたように、個々の時系列データはクモに捕獲される虫の飛行軌跡とみなされる。時系列中の各点$\mathbf{x}_t^k$はｔ時点における虫の位置に対応する。なお「虫の位置」は比喩的な表現であり、アルゴリズム上３次元である必要は無いことに注意されたい。

適応過程に入る以前に、Ｎ個の虫の飛行軌跡は次の式にしたがって標準化される。

$$Z = \frac{X - \mu}{\sigma}$$

ここで、μとσは、全てのデータ点$\mathbf{x}_t^k$の散布図から各次元への射影について計算したｎ組の平均と標準偏差である。これにより単純に各次元から見た分布を平均０、標準偏差１に標準化している。将来的には主成分分析等でより強力な標準化を考えるべきかもしれない。

2.3. クモ

クモは次の４つの特徴を持つように設計した。

- クモはそれぞれ１個の巣を持ち、虫を捕らえる。
- クモは虫を捕えた後に巣の位置を微調整する。
- 虫を捕らえられないクモは捕らえられるようになるまで移動を続ける。
- テリトリーが重なりあうクモは生存競争を行い。勝者のみが生き残る。

クモの巣は超平面と中心点で表される。虫の軌跡が、クモの巣の超平面上の中心点から半径ｒの円内を、角度θ以内で突入したときに、その虫はそのクモの巣に捕獲されたと判定する。ただし「捕獲」は比喩的な表現で、虫の飛行軌跡は変化しない、すなわち捕獲されても虫は失われないものとしたことに注意されたい。

捕獲後の巣の微調整は、単純に中心点を捕獲位置に移動し、虫の飛行経路と直交させるように角度を変化させることで実現した。

クモの移動はランダムウォークである。ランダムウォークの平均歩幅ｍはパラメータとして残した。

クモの生存競争の勝敗は単純に虫の捕獲数の多寡によって決定した。

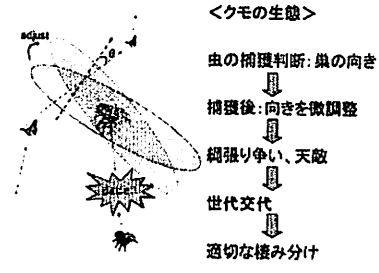


図３ クモの生態

2.4. 世代交代

世代交代の目的は最良の棲み分け状態の実現にある。後のクラスタリングにとって、クモの巣の位置は重要な特徴として使われる。最良の棲み分け状態とは、各クモがそこそこの数の虫を取り合って、データを分離する役に立つ状況をである。前述の生存競争だけで

はすべての虫を取るような一人勝ちするクモが生き残りやすいが、そのようなクモはデータを分離する役にはたたない。何らかの分散圧力が必要である。ここでは、次の手順で現在のクモの巣の状態を評価し、世代交代を導入した。

- 評価中はクモの巣は固定しておく。
- すべての時系列データを入力し、ある虫がどのクモの巣に次々に捕らえられたかの捕獲リストを計算する。
- 虫を一匹でも捕らえたクモを次世代の親の候補とする。
- 候補中でなるべく少ない虫を捕らえるクモを親個体とする。
- 親個体の周辺にランダムに子個体を生成する。

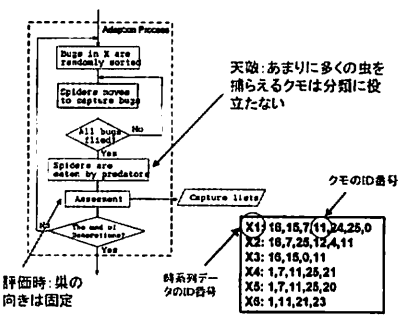
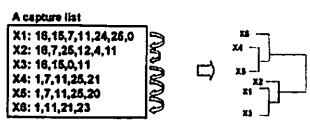


図4 適応過程と捕獲リスト

2.5. 類似度

捕獲リストから類似度を計算するために順序相関〈rank correlation[7]〉を利用した。順序相関とは、連続データではなく、離散データの登場順序のリストから計算される類似度の評価量である。既存の有力な計算方法に「Spearman の ρ」と「Kendall の τ」の2通りがある。詳細は略するが、いずれも完全に同順で一致するとき1、完全に逆順一致するとき－1となる。



■ 順序相関係数

完全一致＝1、完全逆順序＝－1

既存手法

- Spearman's rank correlation coefficient
- Kendall's rank correlation coefficient

図5 順序相関による類似度計算

3. 実験

3.1. データセット

提案手法の評価実験を行った。使用したデータセットはオーストラリアの手話データセット[5]である。発光グローブのビデオ像から作成されたものだが、スケール調整やノイズ除去等の前処理が済んだ便利なものである。come, girl 等１０個の単語について、それぞれ２０個ずつ３次元時系列データが採録されている。その例を図６に示す。

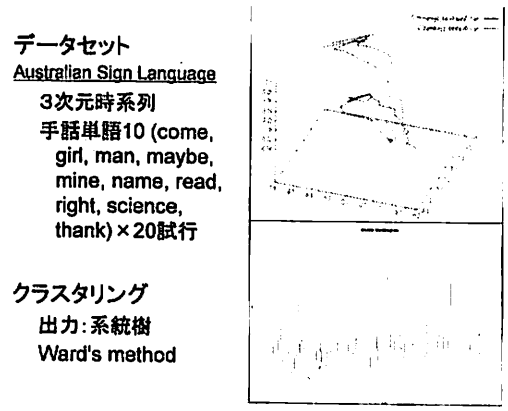


図６ 実験設定

3.2. 評価方法

一般に、クラスタリングの性能を評価することは困難である。クラスタリングはもともと正解の無いデータの分析に用いる教師なし学習だからである。手話データでは幸いにデータには単語を表すラベルがつけられている。そこで、ここではラベルの分類の成否をF-masure[6]に基づいて評価することとした。F-masure は教師付学習の評価のために提案されたもので次のように定義される。

$$F = \frac{2 \cdot \frac{N_1}{N_2} \cdot \frac{N_1}{N_3}}{\frac{N_1}{N_2} + \frac{N_1}{N_3}}$$

ここで、N1 はクラスタ A でラベル L と正しく判定された数、N2 はクラスタ A に含まれるデータ数、N3 はデータ全体でラベル L と判定されるべきデータ数である。F-masure は正答率が高いほど大きく、また感度が高いほど大きい。

クラスタリング方法はスパイダー・アルゴリズムとは独立に選択できる。ここでは、個々のクラスターがどのように成り立っているかの可視化を優先して Ward の方法による系統樹型のクラスタリングを用いた。

3.3. 結果

図７にスパイダーアルゴリズムによる４つの単語（come, name, science, thank）のクラスタリング例を示す。図８に、ユークリッド距離、dynamic time warping(DTW)、スパイダーアルゴリズムの３つをそれぞれ類似度評価に用いたときの、クラスターの分類性能の F-masure をまとて示す。図で、各行は単語を何種類混ぜて提示したかに対応する。例えば、3 signs は、手持ちデータ１０種のうち３種を混ぜて提示したときの分類性能を、１０から３を取る組合せすべてについて計測した平均値である。いずれのケースでもスパイダーアルゴリズムによって計算した類似度を用いた方が、高い F-masure を示している。

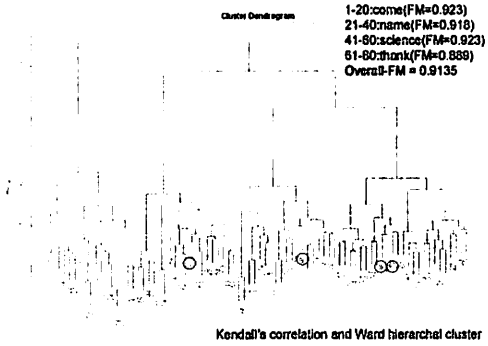


図７ クラスタリング例

	Euclidean distance	DTW	Spider Algorithm
2 signs (${}_{10}C_2$)	0.94	0.94	0.94
3 signs (${}_{10}C_3$)	0.89	0.89	0.90
4 signs (${}_{10}C_4$)	0.85	0.85	0.87
5 signs (${}_{10}C_5$)	0.83	0.81	0.84
6 signs (${}_{10}C_6$)	0.77	0.72	0.82
7 signs (${}_{10}C_7$)	0.78	0.71	0.80
8 signs (${}_{10}C_8$)	0.73	0.67	0.78
9 signs (${}_{10}C_9$)	0.75	0.70	0.77

図８ アルゴリズムの F-masure 比較

4. 結論

時系列データのクラスタリングのための、新しい類似度計算方法として、クモの生態系から啓発されたスパイダーアルゴリズムを提案し、実験的にその性能を調べた。

前節で示した F-masure の良さは「手話」というデータソースの特徴に依存しているかもしれない。手話は相手への伝達を目的とするものであり、個々の単語を区別する特徴が、その軌跡の中に無ければならないからである。その意味で、データマイニング的な使用に耐えられるかどうかは今後の課題である。しかしながら、本研究の発展として、特異値分離テストを実施して、ユークリッド距離等とは全く異なる分類特徴・守備範囲を持つことをすでに確認しており[12]、データマイニングツールとして有効な方法のひとつになりうることを確信している。

文　献

[1] Everitt, B. S.: Cluster Analysis, Edward Arnold, third edition, 1993.

[2] T.Warren Liao: Clustering of time series data – a survey, Pattern Recognition. 38, 2005, pp.1857-1874

[3] Keogh, E. J. and Pazzani, M. J.: Scaling up Dynamic Time Warping for Datamining Application. Principles of Data Mining and Knowledge Discovery 1704, 1999, pp.1-11

[4] Sakoe, H. and Chiba, S.: Dynamic programming algorithm optimization for spken word recognition. IEEE Trans. Acoustics, Speech and Signal Process. Vol.ASSP-26,43-49

[5] Keogh, E and Kasetty, S.: On the Need for Time Series Daat Mining Benchmarks: A Survey and Empirical Demonstration. Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 2002, pp.102-111

[6] Nakamoto, K., Yamada, Y. and Suzuki, E.: Fast Clustering for Time-series Data with Average-time-sequence-vector Generation Based on Dynamic Time Warping. The Japanese Society for Artificial Intelligence 2003, Vol.18 pp.144-152

[7] Keogh, E. and Folias, T.: The UCR Time Series Data Mining Archive Riverside CA. University of California - Computer Science & Engineering Department, 2002, http://www.cs.ucr.edu/~eamonn/TSDMA/index.html

[8] Bjornar, L. and Chinatsu, A.: Fast and Effective Text Mining Using Lineartime Document Clustering. Conference on Knowledge Discovery in Data 1999, pp.16-22

[9] Kendall, M. and Gibbons, J. D.: Rank Correlation Methods, Oxford University Press, fifth edition 1990

[10] Barbar'a,D.: Requirements for Clustering Data Streams,SIGKDD Explorations, 2002 Vol.3, No.2, pp.23-27

[11] Kameda, S. Yamamura, M. : Spider Algorithm for Clustering Time Series, Proc. WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Database (AIKED '06)

[12] Kameda, S. Yamamura, M. : Spider Algorithm for Clustering Multivariate Time Series, WSEAS trans. on Information Science and Applications, 485-492 2006.