1V-04

頻出パターン系列の出現順序に着目したコンテキスト抽出法

大越 寿彦† 渡邊 優太† 太田 昌克‡ 宮崎 敏明†

会津大学コンピュータ理工学部†
日本電信電話株式会社 NTT 未来ねっと研究所‡

1. はじめに

近年、様々なウェアラブル端末やバイタルセンサが開発され、スマートフォンと連携し、ユーザ自身の動作や内部の状態を常時計測できるようになった。常時計測された時系列データには、様々な状況（コンテキスト）におけるデータが含まれている。例えば、人の日常生活においては、「睡眠中」、「歩行中」、「食事中」等の様々なコンテキストが局所的に存在し、それらの変化に応じて、各種センサのデータが変動する。コンテキスト・アウェアなサービスを実現するためには、このような局所的にみられるコンテキストを把握する必要がある。

本稿では、時系列データの特徴的な局所的変化の出現順に着目し、局所的なコンテキストを抽出する手法を提案する。さらに、本手法をウェアラブル心拍計で計測した心拍データに適用した実験結果についても報告する。

2. コンテキスト抽出

時系列データから局所的コンテキストを抽出する手法として、文献[1]の研究がある。これは、時系列データを SAX という文字列変換手法で時系列データを文字列に表現（量子化）した後、頻出する局所的パターンをコンテキストとして抽出する。しかし、当該手法では、出現頻度が高い文字列をパターンとして検出するため、文字列の短いパターンが検出されやすい。従って、短期間のコンテキストは検出されやすいが、幾つかの特定のパターンがまとまって出現する長期のコンテキストを検出することが困難である。

それを解決するために、ここでは、短期間の頻出パターンの出現順序がまとまって出現する頻出パターン群を、長期のコンテキストとして抽出することを試みる。頻出パターンは、文字列で量子化した時系列データから Prefixspan[2]を用いて抽出する。頻出パターンの出現順序を調べるために、頻出パターンをノードとし、連続して出現する頻出パターンに対応したノード同士を、エッジで結んだ無向グラフを作成する。また、エッジにはその出現回数を重みとして付与する。次に、作成した無向グラフをサブグラフ(クラスタ)に分割する。ここでノード間の繋がりが、クラスタ内では密に、クラスタ間では疎となるようなクラスタに分割する。得られた各クラスタ内のノード群は、頻繁に連続して出現する頻出パターン群であることから、各クラスタを１つのコンテキストとして抽出したことになる。

3. 評価実験

心拍データを用いて提案手法の評価を行った。データは筆者自身が図１の心拍計[3]と iPhone アプリ[4]を用いて、ランニング時、安静時、登校時の 3 つのコンテキストにおいて、それぞれ７回ずつ常時計測したデータを用いた。各回の測定時間は平均 21 分で、1 秒おきに計測される心拍データを 10 秒毎に標本化した。頻出パターンの検出に用いたパラメータ値を表１に示す。文字長は 10 秒おきにデータを標本化しているため、1 分間の頻出パターンを抽出している。また、最低出現頻度は 1 つのコンテキストのデータが 7 個あるため、各データに 1 回は出ると考え、7 とした。

図１　心拍計と iPhone アプリ

表１　頻出パターン抽出のパラメータ

量子化の文字種	5
パターンの文字長	6
パターン最低出現頻度	7

SAX によってデータを量子化する際、各量子化文字が示すデータ領域の境界付近において、データの値が少しでも異なると別の文字に量子化されてしまう。このような少しのデータの違い

Context Extracting Method Using Appearing Order of Frequently-Appearing Patterns
†Toshihiko Okoshi, †Yuta Watanabe, ‡Masakatsu Ohta, †Toshiaki Miyazaki
†School of Computer Science and Engineering, The University of Aizu
‡NTT Network Innovation Laboratories, NIPPON TELEGRAPH AND TELLEPHONE CORPORATION

による量子化の揺らぎを吸収するために、階層的クラスタリングを用いて、頻出パターンをグループに分類する。グループ間の距離は群平均法を用いて計算し、グループ内の頻出パターン間の違いが平均 1 文字以内となるグループを求めた。さらに、このグループをノードとし、頻出パターンの出現順序からグループの出現順序を求め、無向グラフを作成した。

グラフのクラスタリングは、次式のモジュラリティが最大となるようにgreedyアルゴリズム[5]を用いて行い、コンテキストを抽出した。

$$Q = \frac{1}{2E}\sum_{i,j}(a_{ij} - \frac{k_i k_j}{2E})\delta(c_i, c_j)$$

ここで、a_{ij} はノード i と j のエッジの重み、E はグラフ全体のエッジの重みの総和、k_i はノード i が持つエッジの重みの総和、c_i はノード i が属するクラスタである。また、$\delta(c_i, c_j)$ はノード i と j が同じクラスタなら 1、違うクラスタなら 0 となる。

接続頻度に閾値を与え、前述のクラスタリングを行った結果を、図 2 及び図 3 に示す。複数のノードすなわち、複数の頻出パターンがまとまって出現するコンテキストが抽出されていることがわかる。ここでは、クラスタ内の頻出パターンの 30%以上が含まれるコンテキストを求め、コンテキスト毎にクラスタを配置している。

図 2 からわかるように、接続頻度 7 以上では、クラスタ 7c を除き、全てのクラスタが「ランニング時」、「登校時」、「安静時」の 3 つのコンテキストのサブコンテキストとなっている。クラスタ 7c は、「登校時」、「安静時」において共通に出現したコンテキストである。これは「登校時」と「安静時」で、心拍データに類似の変動があったことを意味し、2 つの共通する新たなコンテキストともとらえることもできる。

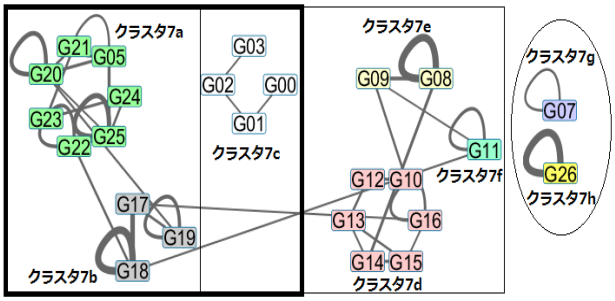

図 2　クラスタリング例 1。接続頻度 7 以上 (Q=0.7358)。四角の太線で囲んだクラスタ群が登校時、四角の細い線が安静時、楕円がランニング時を表す。

接続頻度 2 以上では、接続頻度 7 以上で検出されなかったパターンのグループ（G04、G06）が出現し、さらにクラスタの結合と分割が同一コンテキスト内で一貫して起きている(図 3)。

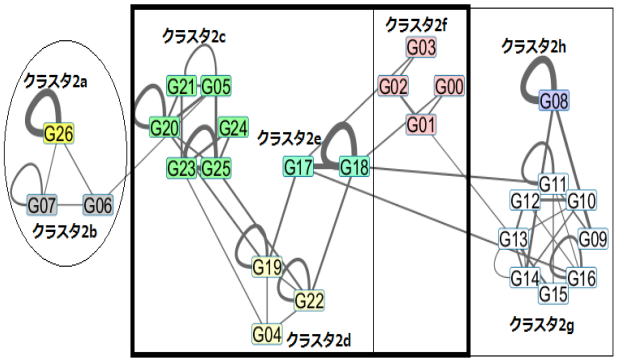

図 3　クラスタリング例 2。接続頻度 2 以上 (Q=0.7227)。四角の太線で囲むクラスタ群が登校時、四角の細い線が安静時、楕円がランニング時を表す。

4. おわりに

時系列データから、頻出するパターンの出現関係を用いて、コンテキストを抽出する手法を提案した。また、心拍データを用いた実験により、頻出パターン系列の接続関係からコンテキスト抽出が可能であることを確認した。本実験では 3 つのコンテキストを想定したが、コンテキストには、人が直感的に気付かないコンテキストも抽出された。本手法はこのような意識することが困難なコンテキストもセンサデータから発見できる可能性がある。今後は、頻出パターン抽出のパラメータや接続頻度閾値の適切な値を検討し、時系列データに含まれる様々なコンテキストを発見していく。

参考文献

[1] Tanaka, Y., Iwamoto, K., & Uehara, K. (2005). Discovery of time-series motif from multi-dimensional data based on MDL principle. *Machine Learning*, *58*(2-3), 269-300.

[2] Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. C. (2001, April). Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In IEEE ICCCN (p. 0215).

[3] PULSENSE: http://www.epson.jp/products/pulsense/ps100/. (accessed 2015/12/18)

[4] Wahoo Fitness : https://itunes.apple.com/jp/app/wahoo-fitness/id391599899?mt=8. (accessed 2015/12/18)

[5] Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. Physical review E, 70(6), 066111.