2B3-02

グラフマイニングを応用した系列データ解析

A Graph Mining Approach for Analysis of Sequential Data

稲積　宏誠[*1]　福田　遼平[*2]　和田　貴久[*2]　大野　博之[*1]
Hiroshige Inazum　Ryohei Fukuca　Takahisa Wada　Hiroyuki Oono

[*1]青山学院大学　理工学部　情報テクノロジー学科
College of Science and Engineering, Aoyama Gakuin University

[*2]青山学院大学大学院　理工学研究科　理工学専攻　知能情報コース
Graduate school of Science and Engineering, Aoyama Gakuin University

This paper presents some aspects of a graph mining approach for analysis of sequential data. One is based on time-scale hierarchy using frequent subtree mining. The other is based on similarity measure using substructure distribution analysis. Both of them provide users with powerful tool set to analyze sequential data.

1. はじめに

グラフ構造のもつ汎用的な表現能力や，理論面および実用面における有用性は広く認知されている．また，グラフ構造をもつデータの取り扱いのひとつとして，グラフマイニングの必要性も広く認識されており，関連研究は広範囲に及んでいる [浅井 04]．ただし，その典型例としては，Web 構造マイニングやリンク解析，また化学構造マイニングのように，明示的なモデルが中心である．

一方，グラフ表現は，命題論理と述語論理の中間クラスとしての知識表現能力があるといわれている [吉田 97]．これは，グラフ構造そのものが明示的に与えられていなくても，あるクラスの知識表現をグラフ構造として表すことが可能であることを意味している．したがって，グラフマイニングアプローチは，広範囲のデータマイニングの課題に対応できるはずである．しかしながら，現実には，さまざまなグラフ構造を扱うことのできるグラフマイニングアルゴリズムが，データマイニングの対象となる問題全般に広く用いられているとはいえない．

グラフマイニングを用いた実問題の解析を困難にしている要因は，大きく分けて 2 つある．一つは，適切なグラフ表現を実現することの困難さであり，どのような問題設定に対してグラフ表現化することが有効なのかという基準や，問題設定に応じたグラフ表現方法の指針構築が求められる．二つ目は何をどのようにマイニングすべきかという明確な基準を設けることの困難さであり，種々のマイニングアルゴリズムに対して，出力内容とその解釈のための環境構築が望まれる．

これまで著者らは，多くの基盤技術 [Nguyen 05] [高林 05] [Zaki 02] を参考，あるいは有効に活用してグラフマイニングの応用に向けた取り組みを行ってきた．[福田 07] [福田 08] [速水 05] [和田 07] [和田 08] 本稿では，これらの取り組みを上記視点から整理することを目的とする．特に，系列データ分析に対するグラフマイニングの適用方法を整理し，木構造表現の有効性と，部分構造の利用を通して類似性の定義，クラスタリングへの応用について述べる，さらにそれらのクラスタリング結果を利用することによって，対象とする系列データの要約情報を構造的に表現できることに言及する．

連絡先: 稲積宏誠，青山学院大学理工学部　情報テクノロジー学科 〒 229-8558 神奈川県相模原市淵野辺 5-10-1 hiro@ina-lab.it.aoyama.ac.jp

2. 問題の所在

本稿では，系列データを取り上げた．たとえば，顧客の購買履歴データは，系列データとして扱うことができる．顧客ごとに，各購買イベントをラベル付きノードとみなし，イベント間の性質的な関係や時間的な関係をラベル付けしたリンクで結合することにより，グラフ表現が可能となる．その結果，購買パターンの分析を行うというアプローチがグラフマイニングの応用領域となる．たしかに，購買履歴データはマーケティング分野で広範囲に活用される対象ではあるが，グラフ表現化されることによってどのような解析上の優位性があるかという点が明確ではないため，その応用例が見られない．したがって，系列データ解析において，まずどのような視点からモデル化を行う必要があるかについて検討しなければならない．

複数の系列データから共通する特徴や傾向を探す場合，細部は完全に一致していなくても，特定の期間などに注目して情報をまとめた場合には共通する傾向が存在する場合や，全体を通じて概要が共通する場合などがある．すなわち，まず系列データに対して，時間情報をどのような単位で評価することができるかということ，単位時間ごとのイベント間の関係をどのように評価することができるかということがモデル化におけるポイントである．さらに，注目する系列データごとの特徴類似性をどのように定義し，評価するかというが分析におけるポイントだといえる．また，このことは，定型的なクエリが設計できない系列データ分析問題に対して，いかなるクエリに基づいて分析することができるのかを発見すること，すなわち多様な表現に基づくクエリを発見することに他ならない．このようにしてクエリが求められるならば，それをもとにして，既存のデータマイニング技術や検索技術を提供することが可能となる．

著者らはまず，系列データに対して，時間単位の階層関係に着目して各イベントの時間情報を多義的に表現することによる系列パターン分析 [中原 05] に注目した．これは，各イベントの時間情報をベクトル表現して分析が行われいるが，実質的に木構造マイニングに相当するとみなすことができる．すなわち，木構造表現された系列データから各ノードを削除した任意の部分木を系列パターンとみなすことである．このとき，中間ノードを削除して得られる木構造表現は，系列表現の一部をワイルドカード化した内容そのものを意味し，その表現を活用することによって，系列パターンの特徴表現の幅が大きく広がる

ことが示されている．このようなアプローチは，部分グラフ抽出ではなく部分木抽出という木構造特有の性質による．

そこで，モデル化の方法としては，時間単位の階層関係を表現する木構造の解析と，汎用的な関係を表現するグラフ構造の解析を実現することとした．また，解析方法としては，部分グラフ（木）抽出を基本とし，各系列データについて，それぞれの含有する部分グラフ（木）の保持パターンを用いた類似性に基づく方法と，各部分グラフ（木）について，それをもつ系列データ群を用いた類似性に基づく方法の 2 とおりの取り組みを行う．以下その概要を示す．

2.1 時間単位の階層関係と頻出部分木の利用 [福田 07]

系列データの局所的，断片的な特徴分析の手法として，次のようなモデル化を行う．

各系列データを時間単位の階層関係に基づいて木構造表現し，それらに共通して包含される部分木を抽出する．ここでの部分木とは，その親子関係あるいは先祖関係のいずれかが対象とする木に共通に含まれているものと定義する．得られた部分木集合を用いて，系列データ集合についての二とおりの要約方法を示す．

1. 各部分木は，それを保有する系列データにより特徴付けられていると解釈する．したがって，系列データ集合の共通性に基づいて類似度を決定し，その類似度に基づいて部分木のクラスタリングを行う．クラスタ内のいかなる部分木も自分自身の部分木とはならないものを基礎構造，そのいずれかの基礎構造となる部分木を包含する部分木を派生構造とし，クラスタ内の部分木集合を基礎構造と派生構造に分類する．その結果，各クラスタの基礎構造と派生構造の組み合わせを表現することによって，系列データ群の特徴を要約表現する．

2. 各系列データは，いかなる部分木集合を包含しているかによって特徴付けられていると解釈する．したがって，部分木集合の共通性に基づいて類似度を決定し，その類似度に基づいて系列データのクラスタリングを行う．各系列データに共通して含まれる部分木集合を共通構造，それ以外の部分木を付加構造に分類する．その結果，各クラスタの共通構造と付加構造の組み合わせを表現することによって，系列データ群の特徴を分類表現する．

先祖関係を許容する部分木は，その最上位の時間単位と最下位の時間単位の間の部分についての関係は特定していないため，その期間内は，ワイルドカードを含む共通パターンとみなすことができる．したがって，いずれの場合も，ワイルドカードを含む頻出系列パターンの組合せで特徴表現が行われる．ただし，前者は，系列データを重複を許してクラスタリングすることによる要約であり，後者は，部分木集合に対して重複を許してクラスタリングすることによる要約を意味している．

2.2 部分構造に基づく構造類似性の利用 [和田 07]

系列データの局所的，断片的な特徴分析ではなく，全体の概要，すなわち大域的な共通性を評価することも重要である．この取り組みは，当初化学物質データを対象として進められ，データ全体の性質を，局所的な情報の組み合わせで評価するための類似性の定義とその算出方法を提案したものである．要約方法としては，前節２つ目の考え方と同様であるが，次の点が異なる．

- 系列データを表現したグラフ構造の各ノードがどの各部分構造に含まれているかという情報を保持する．抽出された部分構造は，どの系列データのどの時点に関係しているのかに基づいて表現される．したがって，各系列データは，部分構造がどの時点にどのように分布しているのかということで特徴付けられることになり，それに基づく類似性が定義され，クラスタリングされることによって系列データ群の特徴を分類表現することができる．

3. おわりに

本稿では，系列データに対してグラフマイニングを用いて要約表現するための考え方を示した．それらは，いずれも部分構造の抽出を基にしたものであるが，木構造表現の特性も活用することで，局所的な特徴抽出と大域的な特徴抽出のいずれの見方も可能であることを紹介した．

こららの考え方に基づく分析ツールはすでに開発を進めており [福田 08][和田 08]，その利用を通して，本稿で述べた考え方に基づく実問題への適用をはかっていきたい．

参考文献

[浅井 04] 浅井達哉, 有村博紀 : 半構造データマイニングにおけるパターン発見技法，電子情報通信学会論文誌，Vol.J87-D1，No.2，pp.79-96 (2004).

[福田 07] 福田遼平，大野博之，稲積宏誠 : 時系列上の階層関係に注目した特徴抽出手法の検討，第 18 回データ工学ワークショップ (2007)

[福田 08] 福田遼平，大野博之，稲積宏誠 : 縮約木抽出法を利用した時間単位の階層関係に基づく系列データ分析システム，第 19 回データ工学ワークショップ，(2008)

[速水 05] 速水 亜希子，他:部分構造の包含関係を指標とするグラフクラスタリングの提案－化学物質を対象として－, 人工知能学会 知識ベースシステム研究会，SIG-KBS-A405, pp.1-6 (2005)

[中原 05] 中原孝信、森田裕之 : ターゲット顧客を識別するためのクレジット購買履歴データを用いたパターン分析，オペレーションズ・リサーチ，Vol51，No.2，pp.89-96 (2006)

[Nguyen 05] Nguyen,P., Ohara,K., Motoda,H., and Washio,T. : Cl-GBI : A Novel Approach for Extracting Typical Patterns from Graph-Structured Data., *Proc.of PAKDD*2005, pp.639-649 (2005)

[高林 05] 高林 健登, 他:グラフ構造データからの特徴的なパターン抽出における探索の効率化, 第 19 回人工知能学会全国大会，2F3-01 (2005).

[和田 07] 和田貴久, 大野博之, 稲積宏誠 : 対象グラフ集合の特性を反映した構造類似性の提案, 日本データベース学会 Letters (DBSJ Letters), Vol.6, No.1, pp.185-188 (2007)

[和田 08] 和田貴久, 大野博之, 稲積宏誠 : 部分構造に基づく構造類似性を用いた特徴抽出システムとその応用, 第 19 回データ工学ワークショップ，(2008)

[吉田 97] 吉田健一, 元田浩 : 逐次ペア拡張に基づく帰納推論, 人工知能学会誌, Vol.12, No.1, pp.58-97 (1997)

[Zaki 02] Zaki,M.J. : Efficiently mining frequent trees in a forest, *Proc.SIGKDD*2002, ACM(2002)