Vol.2014-MPS-98 No.29
Vol.2014-BIO-38 No.29
2014/6/27

順序や生起間隔を考慮したクラスタ系列パターン抽出法の提案

岡田 佳之[1,a)]　福井 健一[2]　沼尾 正行[2]

概要：本研究では, 我々が以前に提案した共起クラスタマイニング（CCM）という手法の改良に取り組む. 共起クラスタマイニングとは, 複数のクラスタ間における共起性とクラスタ内の類似性を同時に考慮し, 共起する 2 つのクラスタの範囲を決定する手法である. しかし，これまでクラスタ内の事象の時系列上の前後関係や生起間隔は簡単のため考慮していない．そこで今回は，新たにそれらを含めたクラスタ系列パターンの抽出を目指す．その中で，ベイズ推定を用いた手法を提案する．

1. はじめに

波形データや位置情報のような，特徴量で表される個々の事象が系列的に発生している事象系列には，頻出する興味深い共起パターンが埋もれていることが多い．共起パターンとは，系列上で互いに近接し，かつ頻出である事象の組み合わせのことを表す．例えば，GPS データにおける停留点間の遷移パターン等が挙げられる．我々は共起パターンを抽出すべく，以前，共起クラスタマイニング（Co-occurring Cluster Mining ;CCM）と呼ぶクラスタリングと頻出パターン抽出を同時に行う方法を提案した．そして CCM を，燃料電池の損傷の際に生じる波形イベントのデータに適用し，構成部材間の力学的な相互作用を抽出した [1]．CCM では，クラスタ内の事象の類似性と共に，時系列上のクラスタ間の共起性も考慮して共起するクラスタのペアを抽出するため，特徴空間においてクラスタリング行った後に，時系列上の共起性を抽出するといった 2 段階法の問題点も解決できる．しかし，既存の CCM では抽出した共起パターンを構成するクラスタ内の事象に対し，時系列上の前後関係や生起間隔は簡単のため考慮に入れていない．本稿では，それらを含めたクラスタ系列パターンの抽出のため，ベイズ推定を用いた新たな CCM を提案する．

1　大阪大学 大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka Uniersity
2　大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University
a)　okada@ai.sanken.osaka-u.ac.jp

2. 共起クラスタマイニング（CCM）

2.1 CCM(既存)

本手法において共起パターンとして抽出する事象の集合 **A**, **B** に対して以下の要件を定める．

- **共起性**: 時系列上で **A** と **B** の事象の出現が近接
- **頻出性**: **A** と **B** の事象が共起する回数が多い
- **類似性**: **A**(**B**) に含まれる事象は十分類似

共起性および頻出性は従来の記号データを対象とする頻出パターンの要件に相当し，類似性はクラスタリングに相当する．上記の全ての要件を満たすクラスタを**共起クラスタ**と呼び，共起クラスタのペアを**共起パターン**と呼ぶ．

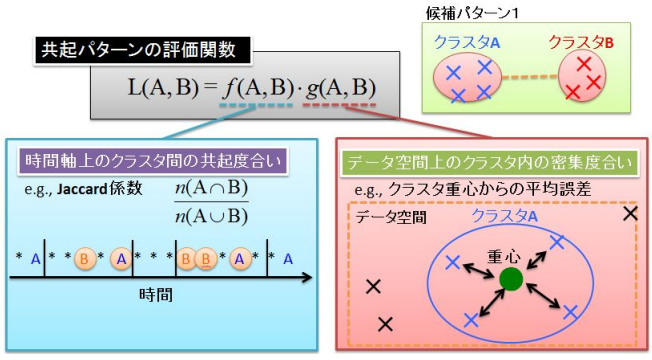

図 1　既存の CCM の概要

具体的なプロセスは，探索空間の縮小のため階層型クラスタリングによって予めクラスタリング過程（デンドログラム）を取得し，その結果得られたクラスタによって全通りのペアを作成する．これらを共起パターンの候補とし，評価関数 L（図 1）が閾値以上の値をとるクラスタのペア

Vol.2014-MPS-98 No.29
Vol.2014-BIO-38 No.29
2014/6/27

を探索する．そして，包含関係にある類似する共起クラスタを除去し，かつ最小支持度以上の頻出性を満たすクラスタのペアを共起パターンとして列挙する．

2.2 順序や生起間隔を考慮した CCM

既存の CCM では得られた共起パターンを構成するクラスタ内の事象の，時系列上の前後関係や生起間隔は考慮していない．これは評価式 L（図 1）において，事象系列を時間の制約によって分割し，分割された区間の中での共起度を Jaccard 係数により測っている（評価式中の f ）ためである．そこで今回は Jaccard 係数を用いることなく，事象間の前後関係や生起間隔を含めた系列パターンの抽出法の提案を目指す．方法として先ず挙げられるのは，階層型クラスタリングによって得られた系列パターンの候補群において，事象間の生起間隔を何らかの閾値に基づいて分割する方法（図 2）である．この場合，その閾値として，事象系列（**A** ⇒ **B**）の生起回数に基づく分割方法が考えられる．しかし，この方法では閾値に対する恣意性が問題となる他，例えば観測事象数が十分でない場合，この方法では統計的に意味のある結果を得ることが難しい．

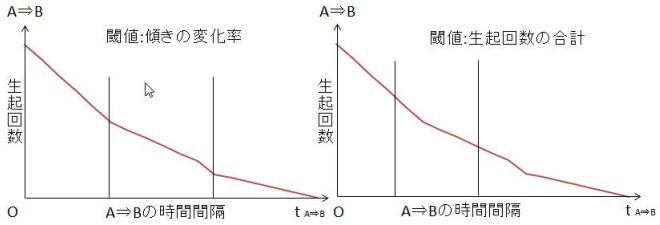

図 2 生起回数の変化率や累積数で分割する場合

そこで，我々はベイズ推定を用いた手法を提案する．ベイズ推定とは，ベイズの定理を用い，観測結果（D）からその原因を規定する母数（θ）の確率を推定するための確率論的方法である．ベイズ推定では，観測結果と，母数に関する既知の情報（事前分布）を用いることで，その母数に関する真の情報（事後分布）を確率分布として推定する．そのため，観測データが少数の場合であっても，頑健に母数に対する真の特性を推定することができる．

今回我々は先ず，時系列上の事象間の前後関係を区別すべく，既存の CCM において階層型クラスタリングを行った結果として現れる共起パターンの候補の選定方法を変更する．既存の CCM における共起パターン候補（**A**,**B**）に対し，今回は **A** と **B** の順序を区別し，**系列パターン候補**を（**A** ⇒ **B**）と（**B** ⇒ **A**）として生成する．次に，事象間の生起間隔を推定すべく，評価式 L においてクラスタ間の時間的な関連度合いを計算する関数 f を図 3 のように定める．まず，それぞれの系列パターン候補に対し，それらを構成する事象 **A**，**B** について生起間隔を測定し，この測定結果から尤度 $\pi(\theta|D)$ を計算する．ここで，θ は事象 **A**,**B**

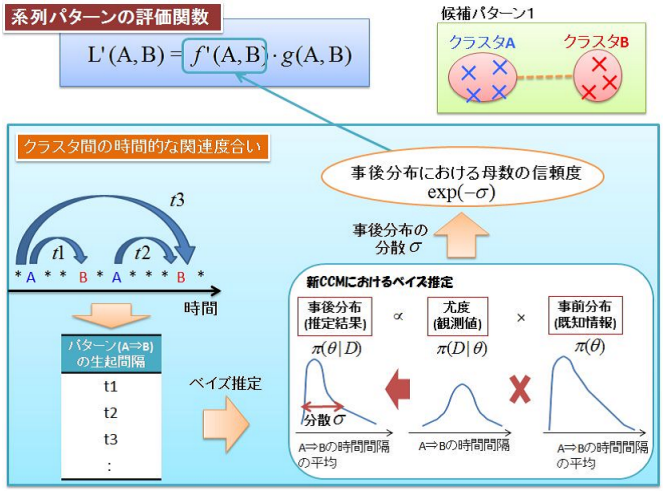

図 3 順序や生起間隔を考慮した CCM の概要

の生起間隔の平均，D は観測データを指す．次に，その観測データに関する既知情報から事前分布 $\pi(\theta)$ を設定し，先の尤度を用いてベイズ推定により事後分布 $\pi(D|\theta)$ を推定する．この事後分布 $\pi(D|\theta)$ の分散値の逆数を評価式中の f に代入する．これは，系列パターンの生起間隔の推定に当たり，ベイズ推定よって得られた推定値が時間的な幅を持たないほど，その推定値における信頼性が高いと考えたためである．これにより，新たな評価式 L のもと，事象間の時系列上の前後関係や生起間隔を考慮したクラスタ系列パターンを抽出できる．なお，抽出したクラスタ系列パターンにおけるそれぞれの事象の生起間隔の分布は MAP 推定により取得が可能である．

3. おわりに

本稿では，特徴量で表される事象の系列において，事象間の時系列上の前後関係や生起間隔を考慮したクラスタ系列パターンの抽出法について述べた．その中で，以前に我々が提案した手法である共起クラスタマイニング（CCM）に対し，ベイズ推定を取り入れた手法を提案した．

謝辞 本研究は JSPS 科研費 24650068 の助成を受けたものです．

参考文献

[1] 稲場大樹, 福井健一, 佐藤一永, 水崎純一郎, 沼尾正行：燃料電池における損傷パターン抽出のための共起クラスタマイニング, 人工知能学会論文誌, Vol. 27, No. 3, pp. 121-132 (2012).

[2] Honda, R. and Konishi, O.: Temporal Rule Discovery for Time-Series Satellite Images and Integration with RDB, in Proc. of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD), pp. 204-215 (2001).

[3] 大澤幸生, 谷内田正彦：キーワード抽出法 KeyGraph の転用による地震履歴データからの要注意活断層発見支援, 人工知能学会誌特集「発見科学」, Vol. 15, No. 4, pp. 665-672 (2000).