A Study on Linkage Pattern Mining in Multiple Sequential Data

その他（別言語等）のタイトル	複数系列データにおける連鎖パタンマイニングに関する研究
著者	Lee Saerom
学位名	博士（工学）
学位の種別	課程博士
報告番号	甲第391号
研究科・専攻	生産情報システム工学系専攻
学位授与年月日	2017-03-23
URL	http://doi.org/10.15118/00009187

氏　　名	イ　セロン 李 セロン
学位論文題目	A Study on Linkage Pattern Mining in Multiple Sequential Data (複数系列データにおける連鎖パタンマイニングに関する研究)
論文審査委員	主　査　准教授　岡　田　吉　史 教　授　板　倉　賢　一 教　授　工　藤　康　生

論文内容の要旨

系列データマイニングは，大規模な系列データから特徴を持つ頻出パタンを発見するデータマイニングの手法である．Agrawal らが 1995 年にその基礎を作り，以後様々なアルゴリズムが開発され，広い分野で利用されている．系列データマイニングにおける多くの研究は，単一または複数の系列データから類似性または相関性をもつ頻出パタンを発見することをねらいとする．

連鎖パタンマイニングは，複数の系列データにまたがって繰り返す頻出パタンの集合（連鎖パタン）を発見する手法である．この手法では，異なる系列データで出現する頻出パタン間で類似性や相関性を示さなくとも，それらが同時刻帯に出現するならば連鎖パタンとして抽出される．連鎖パタンマイニングにより，時系列データに潜む特徴的かつ規則的な変化を示すパタンを発見できるため，バイタルデータや音声データの解析を行う際の有効なツールになると期待される．しかし，既存の連鎖パタンマイニング手法は，ノイズやゆらぎを含む連鎖パタンの抽出精度が極めて低いという問題があった．

本研究の目的は，ノイズに対して頑健な連鎖パタンマイニング手法を開発することである．本手法は，１）系列データの正規化・離散化，２）各系列からの頻出パタン抽出とそれらへのラベル付け，３）ラベル付けされた頻出パタンの区間グラフ生成，４）区間グラフへの飽和集合マイニング，５）連鎖パタン出力，の５ステップで構成される．本手法の新規性は，ステップ４）においてノイズを除去しクリアな連鎖パタンを抽出する点にある．飽和集合マイニングにより，区間グラフ間で共起する極大な頻出パタンのみを得ることができるため，ノイズによって偶然に形成

される偽のパタンを効果的に排除できる．

本研究では，まず，本手法で扱うパラメータのグリッドサーチを行った．結果，頻出パタンにおける最大ウィンドウ幅と最小出現数，飽和集合マイニングにおける最小サポートが小さいほど抽出精度が高いことが示された．続いて，本手法と既存法の性能比較実験を実施した．結果として，本手法により，系列データにおけるノイズの有無に関わらず,連鎖パタンの抽出精度を大幅に向上できることが示された．さらに，本研究では，心電図データへの適用実験を実施した．ここでは，心電図データ特有のピーク値に対処するため，データ分布に基づく離散化法が新たに開発・導入された．結果，心疾患の異常特定に利用される波形からなる連鎖パタンが得られた．これは，連鎖パタンマイニングが心電図データからの新たな異常検出手法として利用できる可能性を示唆している．

ABSTRACT

Sequential pattern mining is a promising and effective data mining method for finding frequent patterns in large-scale sequential data. After Agrawal et al. constructed the foundations of sequential pattern mining in 1995, various new effective algorithms have been developed and applied in a wide range of fields. These research aims to detect same or similar subsequences within a single sequential data or among multiple sequential data.

Linkage pattern mining is a data mining technique that finds frequent patterns that appear repeatedly across multiple sequential data. Even if frequent patterns occurring in the respective sequential data do not show similarity to each other, the set of those patterns is extracted as a linkage pattern if it appears continually within the same period. Linkage pattern mining is expected to become a useful approach in various fields such as vital data monitoring or voice analysis. However, the existing method has an issue that it can hardly extract linkage patterns in sequential data with noise/fluctuations.

The aim of this study is to propose a new noise-robust linkage pattern mining method. The procedure of the proposed method is composed of the following five steps: 1) Normalization and discretization of sequential data, 2) Extracting and labeling frequent patterns from each sequence, 3) generating interval graphs depending on overlapping labels on the time axis, 4) closed itemset mining from the generated interval graphs, 5) outputting the linkage pattern. The main contribution of this study is to extract clear

linkage patterns by excluding noise in Step 4). Closed itemset mining can find maximal frequent pattern that co-occurs among the interval graph, and hence pseudo patterns accidentally constructed by noise can be excluded effectively.

In this study, we first conduct a grid search for the parameter values of the proposed method, *maximum window width* and *minimum number of occurrences* in frequent pattern mining and *minimum support* in closed itemset mining. As a result, it is shown that these parameter values should be set to small values for high extraction accuracy. Subsequently, performance comparison between the proposed method and the previous method is conducted using artificial sequential datasets. By this experiment, it is shown that the proposed method can significantly improve extraction accuracy of linkage pattern in sequential data with noise as well as those without noise. Furthermore, the proposed method is applied to real ECG (electrocardiogram) data, and the performance is evaluated. In this experiment, a discretization method based on data distribution is newly incorporated into the proposed method in order to deal with the peak in ECG data. As a result, it is shown that the proposed method can extract meaningful linkage patterns that are composed of waves crucial for diagnosis of heart disease. This suggests that the proposed method is available as a new abnormality detector for ECG data.

論文審査結果の要旨

連鎖パタンマイニングとは，複数の系列データにまたがって連鎖的に繰り返す頻出パタンの集合（連鎖パタン）を発見する技術である．既存の手法では，ノイズを含む時系列データでは正しく連鎖パタンを抽出できないという問題があった．

提出論文は，ノイズに対して頑健な新しい連鎖パタンマイニング手法を提案するものである．提案手法は，１）系列データの正規化・離散化，２）各系列からの頻出パタン抽出とそれらへのラベル付け，３）ラベル付けされた頻出パタンの区間グラフ生成，４）区間グラフへの飽和集合マイニング，５）連鎖パタン出力，の５ステップで構成される．提案手法の最大の利点は，ステップ４)においてノイズを除去しクリアな連鎖パタンを抽出する点にある．飽和集合マイニングにより，区間グラフ間で共起する極大な頻出パタンのみを得ることができるため，ノイズによって偶然に形成される偽のパタンを効

果的に排除できる．

本論文では，まず，人工データを用いた性能評価実験を行っている．ここでは，パラメータのグリッドサーチをとおして各パラメータが抽出精度に与える影響を明らかにし,さらに系列データにおけるノイズの有無に関わらず,本手法の抽出精度が既存法のそれを大幅に上回ることを示している．また，計算時間の計測実験により，本手法で新たに実装された飽和集合マイニングは小規模な区間グラフデータを対象とするため計算負荷の増加は少なく，総合的な計算時間においても既存法とほぼ変わらないことを示している．

次に，本論文では，本手法を心電図データに適用した結果について記述している．ここではまず，心電図データ特有のピーク値に対処するため，データ分布に基づく離散化法を新たに開発・導入している．実験として，心筋梗塞の心電図データにおける異常波形検出を行っている．結果，新たに導入された離散化法により異常波形の検出精度が向上すること，および，複数部位にまたがる心筋梗塞特有の異常波形（異常Q波，ST上昇）を検出できることが示されている．

以上の成果は，生産情報システム工学の中でも特に系列データマイニング分野の発展に寄与するものであり，また，実データに対する高い応用可能性を示すものである．よって，本論文は博士（工学）の学位論文に値すると認める．