様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 １４ 日現在

機関番号：１３９０３
研究種目：若手研究(A)
研究期間：2017～2020
課題番号：１７Ｈ０４６９４

研究課題名（和文）パターンマイニングと疎性モデリングに基づく大規模系列データからの知識創出

研究課題名（英文）Knowledge extraction from large-scale sequence data by combining pattern mining and sparse modeling

研究代表者

烏山　昌幸（Karasuyama, Masayuki）

名古屋工業大学・工学（系）研究科（研究院）・准教授

研究者番号：４０６２８６４０

交付決定額（研究期間全体）：（直接経費）　　16,600,000円

研究成果の概要（和文）：多様なセンサーデバイスの発展やストレージの大容量化に伴い，収集したデータから統計的な推論によって有益な知見を引き出すデータ駆動型の解析の重要性が高まっている．本課題では，系列や，より複雑な接続関係を持つ構造であるグラフで表現されたデータから，解釈可能な重要部分構造を発見するための機械学習アルゴリズムの構築を行なってきた．あり得る部分構造は組み合わせ的に存在するが，本研究では不要なパターンの枝刈り規則を最適化理論に基づき構築することで，最適性を失うことなく重要な部分構造が発見できることを示した．

研究成果の学術的意義や社会的意義
機械学習技術の注目が高まるつれ，その解釈性は大きな社会的関心事となっている．本課題で扱うような系列やグラフは，時系列データや化合物などの科学データで単純な数値テーブルでは表現できないデータの表現方法として広く定着しているものである．そのため，多様化するデータ駆動解析において，本課題で扱ったような問題設定は今後ますます顕在化すると考えられる．一方で，組み合わせ的に爆発する部分構造に対し，最適性を保証しつつ学習する枠組みはほとんど研究がなく，本研究の独自性・意義を示すものである．

研究成果の概要（英文）：Because of development of a variety of sensor devices and increase of the capacity of storage devices, a significance of data-driven approaches to extracting useful knowledge from accumulated data have been widely recognized. In this study, we consider identifying important substructures in sequence or more complicated graph data, which has a connected structure inside a data instance. Since combinatorially many possible substructures exist, we propose efficient pruning algorithms that removes unnecessary patterns. Our framework is based on an optimization theory, and thus, the optimality of the resulting identified substructures can be guaranteed.

研究分野：機械学習

キーワード：機械学習　疎性モデリング　パターンマイニング　構造データ

１．研究開始当初の背景

多様な計測機器の発達に伴って，収集したデータから知見を引き出すビッグデータ利活用の重要性が広く認知されるようになってきた．特に，カメラや GPS センサー，ビーコン等による継時的な測定や DNA 配列やテキスト情報など「点」ではなく「系列」として意味を持つデータや，また，化学や創薬の分野における分子構造，材料科学における結晶構造のデータなど，より一般性のある関係性表現である「グラフ」を用いて整理されるような，より複雑な構造を持つデータも数多く扱われるようになっている．このようにデータの多様性が増したことで，蓄積された情報から意味のあるものを抽出し，解釈することは非常に難しくなっている．機械学習では与えられたデータから何らかの予測を行うモデルを推定することが多いが，機械学習の意思決定に対する解釈性が近年では非常に重要視されている．系列やグラフのような構造を持つデータの場合，予測にとってどこが重要な部分構造（パターン）なのかを同定することは，部分構造が持つ組み合わせ的性質のために非常に難しい．しかしながら，この問題に対する先行研究はそれほど行われていないのが現状である．例えば，本課題と関連の深い[1]はブースティングのアイデアに基づいて系列データからのモデル構築に取り組んだが、パターンを一つづつ追加するこのアプローチのスケーラビリティは高いとは言い難く，多様な構造データへの対応は難しい．

２．研究の目的

予測モデルの推定において，重要な変数を発見するためのアプローチに疎性（スパース）モデリングがある．ただし，この方法を用いても，膨大な候補パターンを扱うことは計算量的に困難性が高い．本研究では，疎性モデリングの最適化問題としての性質を利用し，膨大な候補パターンから不要なものをあらかじめ除去するスクリーニング法の開発を行う．先行研究として，疎性モデリングの不要な変数を除去するセーフスクリーニング[2]が知られていたが，この方法でも，そもそものスクリーニング対象が列挙不可能なほど膨大なケースへの適用は難しい．系列やグラフを持つデータを効率的に数え上げる手法としてパターンマイニングがよく用いられる．パターンマイニングではパターンの出現頻度に単調性ができるように探索木を作ることで頻度の低いものの枝刈りを実現している．この考え方を単純な頻度からセーフスクリーニングの考え方に拡張し，列挙中に，予測に対する重要性で枝刈りを行う枠組みを提案する．これにより，全ての可能性を列挙することなく不要パターンのスクリーニングが可能であり，かつセーフスクリーニングに基づくため結果の最終的な最適性を保証することも可能になる．本研究の目的は，このアプローチによって，多様な構造データの解析（特に予測モデル推定）において，解釈可能な部分構造を同定できる機械学習の枠組みを構築することとなる．

３．研究の方法

疎性モデリングとして最も有名な手法に LASSO(least-absolute shrinkage and selection operator)が知られている．この枠組みは線形モデルの変数選択を凸最適化問題に帰着させる方法であり，組み合わせ的な変数選択問題を回帰係数に対するペナルティ付きの誤差最小化で表現することで最適解を安定的に得ることができる．LASSO は汎用性が高く，様々な分野で非常に広く利用されているため,本研究でもこの手法をベースとして考えていく.申請者はこれまでに,構造データからの LASSO モデル構築のためのセーフスクリーニングの拡張と効率化に取り組んでいた[3]．部分構造の出現の有無を変数とし，LASSO で重要な変数を選択することで最適性保証のある部分構造選択が可能になる．本研究では，このアプローチを方法論的な土台とし，距離学習など発展的な設定への拡張や，連続値データを含むより複雑な構造データの取り扱い，単調性の仮定の緩和，応用問題への適用などに取り組んできた．以下で，それぞれの具体的な結果について述べていく．

４．研究成果

4-1）連続値属性付き系列/グラフデータからの予測マイニング

系列やグラフデータには連続値の属性情報が付くことが多い．例えば，時系列の座標情報や化合物の元素に付随する物理量などがこれに該当する．このようなケースでは従来の方法で特徴的な部分構造パターンを発見することは容易ではなかった．連続値の情報から特徴的なパターンを抽出するアプローチを大別すると，空間中に代表パターンを定めるか，空間を分割してどのような部分空間に属するかでパターンを表現するかがある．空間中に代表パターンを定める場合はどのように代表を決めるかが問題となるが，実践上はクラスタリングなどで決めるか，あるいは代表パターン自体を最適化する．しかしながら，いずれのアプローチも，得られた代表点の最適性を確保することが難しくなってしまう．また，典型的には代表点との近さをスコアにして特徴量を作るが,このスコアの解釈がしにくいケースもある.一方で,空間分割のアプローチでは,空間中で範囲を設けて，その範囲に属性値が入ったかどうかで特徴量を作る．この方法は解釈性

が非常に高いが，あり得る空間の分割が膨大になってしまうという問題があった．
そこで，[4,5]では，空間を分割し数え上げるマイニング技術を，系列やグラフの探索木に組み込んだ新たな部分構造ベースの予測モデル学習法を提案した．例えば，系列データでは系列を拡張する操作と，系列の各点における属性値空間の分割を交互に行うような探索木を構築する．このようにすることであり得る部分構造(系列やグラフと属性値の空間分割の組み合わせ)を網羅できることを示した．また，この探索木上で適用できるセーフスクリーニングによる枝刈りを導出し，最適性を保証しつつ，予測に寄与する部分構造を効率的に検出できる枠組みを構築した．

4-2) 距離学習における予測マイニング
入力データから何らかの対象に対して，直接予測値を与えるモデルの学習に対して，入力データ間の関係性（類似性，距離）を学習する距離学習の重要性が近年増している．例えば，顔認識のモデルでは，認証された顔と登録されているユーザーの顔との類似性を精度良く測る必要がある．あるいは，類似化合物の検索システムの構築を考えた場合に，過去のデータにおいて似た性質を持っていた化合物間の類似度を高くするような推定問題が考えられる．このようなケースでも，類似性の判断にはどの部分構造が重要なのかを明確にできれば明らかに有用である．
そこで，[6-8]では，距離学習の設定においてセーフスクリーニングの理論を拡張し，その後，その知見をもとに[9-11]では，グラフデータに対して，組み合わせ的に存在するパターンを列挙しつつ,最適な距離関数に必要のないパターンを枝刈りする最適化法を構築した．この方法では，距離学習によって分類に適した空間が獲得でき，かつその空間はどの部分グラフに対応するかを明示的に知ることができる．論文では提案法がその解釈性の高さにかかわらず，深層学習によるグラフ分類と同程度の精度を達成するケースがあることを示した．

4-3) 単調性仮定の緩和
ここまで議論してきた方法では，パターンの探索木に対して特徴量が単調に変化することを仮定していた．この仮定は自然な仮定ではあるものの，常に成り立つとは限らないものであった．そこで，[12]では，単調性の代わりにパターンの最大サイズを使った枝刈り基準を導出した．これにより，これまでの枠組みでは扱えなかったジャッカード係数など単調性のない特徴表現にも予測マイニングにより，最適性を保証しつつ多様なパターンを探索することが可能になった．

4-4) 応用研究
応用研究として,動物移動データの軌跡解析[13]や,移動データにおける群間差の検定[14-15],タンパク質配列解析[16-17]の研究を行なった．[13]では動物の移動データに対して系列データ版の予測パターンにマイニングを行い解析を行なっている．[14-15]では，軌跡データにおける群間の差の検定に枝刈りを導入することで効率化する方法論を導入している．[16-17]では，タンパク質合成のための配列の予測に系列ベースの回帰モデルを用いている(ここでは，高次のパターンのない単純なモデルではあるが，機械学習の予測に基づき実験の意思決定に応用している点で意義深く，より複雑な組み合わせ効果までの検討は今後の検討となる)．

[1] H. Saigo, et al., Machine Learning, 75(1): 69-89, 2009.
[2] L. E. Ghaoui, et al., arXiv:1009.4219, 2010.
[3] K. Nakagawa,et al., ACM SIGKDD, pp.17851794, 2016.
[4] 朝日他，情報論的学習理論と機械学習研究会(IBISML)，信学技報，to appear.
[5] 柴原他，情報論的学習理論と機械学習研究会(IBISML)，信学技報，vol.119，no.89，pp.57-64，2019年6月.
[6] T. Yoshida, et al., Neural Computation, vol.31, no.12, pp.2432-2491, 2019.
[7] T. Yoshida, et al., Proc. 24th ACM SIGKDD, pp. 2653-2662, 2018.
[8] 吉田他，情報論的学習理論と機械学習研究会(IBISML)，信学技報，vol.117，no.293，pp.219-226，2017年11月.
[9] T. Yoshida, et al., Machine Learning, to appear.
[10] T. Yoshida, et al., Proc. 25th ACM SIGKDD, pp. 1026-1036, 2019.
[11] 吉田他，情報論的学習理論と機械学習研究会(IBISML)，信学技報，vol.118，no.284，pp.151-158，2018年11月.
[12] 羽川他，ニューロコンピューティング研究会（NC），信学技報，vol. 120，no. 403，pp. 174-179，2021年3月.
[13] T. Sakuma, et al., Advanced Robotics, vol.33, no.3-4, pp.134-152, 2019.
[14] D. N. L. Vo et al., IEEE Transactions on Knowledge and Data Engineering, 2020 (Early Access).
[15] D. N. L. Vo et al., Proc. 27th ACM SIGSPATIAL, pp. 548-551, 2019.
[16] K. Inoue, et al., Communications Biology 4, 362, 2021.
[17] M. Karasuyama, et al., Scientific Reports 8, 15580, 2018.

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　7件／うち国際共著　1件／うちオープンアクセス　3件）

１．著者名 V. N. L. Duy, T. Sakuma, T. Ishiyama, H. Toda, K. Arai, M. Karasuyama, Y. Okubo, M. Sunaga, H. Hanada, Y. Tabei, I. Takeuchi	４．巻 -
２．論文標題 Stat-DSM: Statistically Discriminative Sub-trajectory Mining with Multiple Testing Correction	５．発行年 2020年
３．雑誌名 IEEE Transactions on Knowledge and Data Engineering	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TKDE.2020.2994344	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 T. Yoshida, I. Takeuchi, and M. Karasuyama,	４．巻 31
２．論文標題 Safe Triplet Screening for Distance Metric Learning	５．発行年 2019年
３．雑誌名 Neural Computation	６．最初と最後の頁 2432-2491
掲載論文のDOI（デジタルオブジェクト識別子） 10.1162/neco_a_01240	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Sakuma Takuto、Nishi Kazuya、Kishimoto Kaoru、Nakagawa Kazuya、Karasuyama Masayuki、Umezu Yuta、Kajioka Shinsuke、Yamazaki Shuhei J.、Kimura Koutarou D.、Matsumoto Sakiko、Yoda Ken、Fukutomi Matasaburo、Shidara Hisashi、Ogawa Hiroto、Takeuchi Ichiro	４．巻 33
２．論文標題 Efficient learning algorithm for sparse subsequence pattern-based classification and applications to comparative animal trajectory data analysis	５．発行年 2019年
３．雑誌名 Advanced Robotics	６．最初と最後の頁 134～152
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/01691864.2019.1571438	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 K. Kanamori, K. Toyoura, J. Honda, K. Hattori, A. Seko, M. Karasuyama, K. Shitara, M. Shiga, A. Kuwabara, and I. Takeuchi	４．巻 97
２．論文標題 Exploring a potential energy surface by machine learning for characterizing atomic transport	５．発行年 2018年
３．雑誌名 Physical Review B	６．最初と最後の頁 125124
掲載論文のDOI（デジタルオブジェクト識別子） 10.1103/PhysRevB.97.125124	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 T. Yoshida, I. Takeuchi, and M. Karasuyama	4．巻 -
2．論文標題 Distance Metric Learning for Graph Structured Data	5．発行年 2021年
3．雑誌名 Machine Learning	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 K. Inoue, M. Karasuyama, R. Nakamura, M Konno, D. Yamada, K. Mannen, T. Nagata, Y. Inatsu, H. Yawo, K. Yura, O. Beja, H. Kandori, and I. Takeuchi	4．巻 4
2．論文標題 Exploration of natural red-shifted rhodopsins using a machine learning-based Bayesian experimental design	5．発行年 2021年
3．雑誌名 Communications Biology	6．最初と最後の頁 362
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s42003-021-01878-9	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 M. Karasuyama, K. Inoue, R. Nakamura, H. Kandori, and I. Takeuchi	4．巻 8
2．論文標題 Understanding Colour Tuning Rules and Predicting Absorption Wavelengths of Microbial Rhodopsins by Data-Driven Machine-Learning Approach	5．発行年 2018年
3．雑誌名 Scientific Reports	6．最初と最後の頁 15580
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-018-33984-w	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計10件（うち招待講演　0件／うち国際学会　4件）

1．発表者名 D. V. N. Le, T. Sakuma, T. Ishiyama, H. Toda, K. Arai, M. Karasuyama, Y. Okubo, M. Sunaga, Y. Tabei, I. Takeuchi
2．発表標題 Statistically Discriminative Sub-trajectory Mining with Multiple Testing Correction
3．学会等名 the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems（国際学会）
4．発表年 2019年

１．発表者名
T. Yoshida, I. Takeuchi, and M. Karasuyama

２．発表標題
Learning Interpretable Metric between Graphs: Convex Formulation and Computation with Graph Mining

３．学会等名
The 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining（国際学会）

４．発表年
2019年

１．発表者名
T. Yoshida, I. Takeuchi, and M. Karasuyama

２．発表標題
Safe Triplet Screening for Distance Metric Learning

３．学会等名
The 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining（国際学会）

４．発表年
2018年

１．発表者名
吉田知貴，竹内一郎，烏山昌幸

２．発表標題
部分グラフに基づくグラフ間の距離学習

３．学会等名
情報論的学習理論と機械学習研究会

４．発表年
2018年

１．発表者名
M. Karasuyama and H. Mamitsuka

２．発表標題
Factor Analysis on a Graph

３．学会等名
International Conference on Artificial Intelligence and Statistics（国際学会）

４．発表年
2018年

１．発表者名
吉田知貴，竹内一郎，烏山昌幸

２．発表標題
マージン最大化距離学習におけるセーフスクリーニング

３．学会等名
情報論的学習理論と機械学習研究会

４．発表年
2017年

１．発表者名
烏山昌幸，竹内一郎

２．発表標題
系列データからのクラス特異的代表パターン選出：分類モデルとMorse Complexによるアプローチ

３．学会等名
情報論的学習理論と機械学習研究会

４．発表年
2017年

１．発表者名
朝日陽向，烏山昌幸

２．発表標題
属性区間付きグラフを用いた予測グラフマイニング

３．学会等名
情報論的学習理論と機械学習研究会

４．発表年
2021年

１．発表者名
羽川晟史，烏山昌幸

２．発表標題
予測パターンマイニングにおける非単調性特徴量のためのSafe Pattern Pruning

３．学会等名
ニューロコンピューティング研究会

４．発表年
2021年

１．発表者名 柴原芳和，佐久間拓人，竹内一郎，烏山昌幸
２．発表標題 適応的空間分割に基づく連続値時系列データのためのPredictive Sequence Mining
３．学会等名 情報論的学習理論と機械学習研究会
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関