様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 ２２ 日現在

機関番号：１７２０１
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００２７９

研究課題名（和文）パレート学習型自己組織化マップのマルチモーダル，大規模データ解析への応用

研究課題名（英文）Application of Pareto learning SOM to multi-modal big data analysis

研究代表者
堂薗　浩（Hirtoshi, Dozono）

佐賀大学・工学（系）研究科（研究院）・准教授

研究者番号：００２１７６１３

交付決定額（研究期間全体）：（直接経費）　　3,900,000円

研究成果の概要（和文）：本研究では，研究者がマルチモーダルデータの解析のために開発したパレート型自己組織化マップを，ゲノム解析，パケット解析，講義履修者の履修状況の解析，株価等の時系列データの解析に応用する手法について研究を行った．研究としてはゲノム解析への応用を主なものとし，大規模ゲノムデータから，これまで解析で用いられた頻度情報に加え，より細かいコンテキスト情報や，塩基間の相関係数をマルチモーダルな特徴量用い，パレート型学習SOMで解析を行った．また，パケット解析に関しては大規模データ解析に有用であると考えられるCGH-SOMを用いた解析法を考案した．

研究成果の概要（英文）：In this research,we applied the Pareto type Self organizing Maps, which are developed for the analyses of multi-modal data to genome analysis, IP-Packet analysis, the analysis of students in the lecture class and time series such as stock price.Mainly, the genome analysis was conducted, and form big data of genome data, the context and correlation coefficients among the nucleotides are calculated with the frequencies of nucleotides, and they are analyzed as multi-modal features by pareto type learning SOM. As for IP-packet analysis, we applied CGH-SOM which is effective for big data analysis.

研究分野：ソフトコンピューティング

キーワード：マルチモーダルデータ解析　ビッグデータ解析　自己組織化マップ　メタゲノム解析　IPパケット解析

１．研究開始当初の背景

パレート学習型自己組織化マップは本研究費の申請者が 2008 年に発表し、その有用性が示された自己組織化マップの１つである。自己組織化マップ(SOM)は Kohonen により提案されたニューラルネットワークモデルで，入力層に与えられた多次元ベクトルを教師なし学習し，入力ベクトル間の関係を出力層上に自己組織化する。このような特性から，多次元ベクトルのクラスタリングおよびベクトル間の関係の２次元平面上での可視化などに用いられている。これまで申請者の研究グループでは，染色体プロフィールデータのクラスタリング，DNA シーケンスの解析，バイオメトリクス認証、自律移動ロボットの制御、IP パケットのクラスタリングなどへの自己組織化マップの応用を行ってきた。これらの研究の中で、バイオメトリクス認証、特に、行動的特徴量を用いるバイオメトリクス認証において、特徴量の解析、および、認証システムの構築のためにパレート型自己組織化マップは提案された。

行動的特徴量は，キー入力の癖，マウスの動作パターン，ペンの動き等パソコン等に装備されたデバイスで取得可能で，複数の行動的特徴量を統合してマルチモーダルな認証を行う方法を申請者は提案してきた．

このマルチモーダルな特徴量の統合を行う　ために，複数のベクトルをそれぞれ独立のベクトルとして、多目的最適化問題におけるパレート最適性の考え方を用いて学習を行う自己組織化マップとして、パレート学習型自己組織化マップ(Pareto learning SOM: P-SOM)を開発した．

２．研究の目的

これまでの研究をふまえて，本研究の第１の目標は、マルチモーダルなデータ解析への応用である。前述のように申請者は、生体認証、DNA シーケンス解析、自律移動ロボットの制御、IP パケットのクラスタリングなどに自己組織化マップを用いてきた。生体認証問題に対してはこれまでもマルチモーダルな特徴量に対して，SP-SOM を適用してきたが、さらに様々な生体情報の統合、および、リアルタイムでのユーザ識別のためのパソコンの使用時のオンライン学習および認証への応用を進めていく。

また，SOM のニューロンはベクトルをニューロン値として保持しているが、ニューロン上の値は必ずしもベクトルである必要はなく、入力ベクトルとの距離が定義できるものであればよく、ニューラルネットをニューロン上に学習するものや、隠れマルコフモデルを学習するものも存在する。SP-SOM, P-SOM においては隠れマルコフモデルとベクトルの組み合わせや、複数の隠れマルコフモデルの組み合わせ等も学習可能であると考えられる．このような時系列モデルの組み合わせに関しては，次に述べる IP パケットの解析や DNA シーケンスの解析に有用であると考えられる．

本研究の２番目の目標は，大規模データの解析である．大規模データの解析としては，１昨年度より IP パケットトラフィックの解析の問題にも SP-SOM, P-SOM を用い，SOM を用いた場合と同様に，パケットトラフィックの時間的な変化を反映したマップを作成することに加え，SP-SOM を用いた教師付き学習による不正トラフィックの検出，および，P-SOM を用いた教師無し学習による不正トラフィックの検出が可能であることを示した．本研究ではより解析精度を上げるため，通常は画像データや３次元オブジェクトデータの解析に有用な CGH-SOM の応用を考えている．すなわち，従来の SOM などを用いて，パケットトラフィックを２次元データとして可視化を行い，そのデータを CGH-SOM を用いることで，パケットトラフィックの解析を行い，異常検出を行う方法を考えている．

また，大規模データ解析として DNA シーケンスの解析を行いたいと考えている．近年，次世代シーケンサの登場により，大量のシーケンスデータが生成され，その解析方法が問題となっている．申請者は以前 DNA シーケンスの解析および DNA チップの設計手法として自己組織化マップを用いた方法を提案したが，以前用いた方法は DNA を文字列として学習する方式で，大規模なデータ解析には向かないと考えられる．これに対し，特定長(3-5 程度)の全てのシーケンスの出現頻度をカウントして入力ベクトルを作成し，SOM に学習させることで，生物種ごとの特徴や同じ生物種内での進化に合わせてマッピングされることが示されており，申請者のグループでも WEB 上から自動的にデータを取得し，この方式で学習を行う実験を行い，有効性を確認した（図３参照）．この出現頻度をカウントする方式は，大規模データの解析には有効であり，前述の IP パケットの解析にも用いられている．本研究ではこの方式に対して，複数長のシーケンスの出現頻度を用いた場合，また，DNA シーケンスの文脈情報を加えた場合，DNA シーケンスの隠れマルコフモデルを加えた場合など，マルチモーダルな入力ベクトルを用いて，より精度の高いマッピングを可能とし，シーケンスデータから生物種やーケンスの種別などを判別できるシステムを構築する．

３．研究の方法

・SOM，P-SOM，SP-SOM のよりマルチモーダルな学習システムへの拡張

SOM< P-SOM,SP-SOM のニューロンにベクトル値のみでなく、隠れマルコフモデルやニューラルネットなどの構造を組み込み、ベクトル値と組み合わせたり、複数の構造を組み合わせて用いることで、より表現能力の高い自己

組織化マップを開発する。

・ パケットトラフィック学習アルゴリズムの開発

これまでに引き続き，パケットキャプチャソフトを用いて収集した IP パケットトラフィックを学習するアルゴリズムを開発する。この際，入力データとしてパケットのヘッダやパケット内のデータをベクトル化したものを用いる。ヘッダに関しては，どの要素をまとめてベクトル化し，また，逆にどの要素を分割してベクトル化するかが問題となり，数学的な解析または実験的な解析が必要となると考えられる。また，パケット内のデータに関してはサイズが大きいため，そのまま自己組織化マップに学習させることは困難であると考えられる。そのため，なんらかの統計的な情報を用いるか，1 バイトコード(00H-FFH)の発生頻度を用いるなどのデータの変換が必要になると考えられる。

・マルチモーダル生体認証システムへの応用

これまでの研究(研究業績(3)他)に引き続き、個人認証システムへの応用を進めていく。ここではキーボード，タブレット、マウスやゲームコントローラなど様々なデバイスを用いて、パソコンのみでなくゲーム機や、携帯電話、iPOD などでも利用可能なマルチモーダル行動的特徴量に基づく生体認証方式の開発を行う。さらに、自己組織化マップの汎化能力や学習能力を生かし、行動的特徴量にみられる入力の変動やユーザの慣れによる認証データの時間的変化に追従可能な認証システムを開発する。

・ DNA シーケンスの学習への応用

WEB 上のデータベースなどから取得した DNA シーケンスを用いて，複数の系列長の出現頻度，文脈情報などをマルチモーダルに組み合わせてパレート型学習 SOM を用いて学習し，従来の単一系列長での学習結果の比較する．

・ CGH-SOM の基礎研究

CGH-SOM は今年度から研究を始め，現状では基礎的な研究の状況にあるため，本格的にマルチモーダル化を行う前に，ホログラムのマッチングの評価方法，ホログラム計算処理の高速化，計測データのホログラム化の前処理の方法などの基礎研究を進めておく．

・ CGH-SOM による IP パケットトラフィック学習アルゴリズムの開発

前述の頻度情報を用いた IP パケット解析法の発展系として，CGH-SOM を用いたパケット解析手法について研究を行う．CGH-SOM は元２次元画像情報や３次元物体の情報のクラスタリングや分類のために開発された SOM であるが，CGH の特性として，変化に敏感であり，より感度の良いパケット解析システムが作成可能と考えられる．そこでパケットトラフィックデータを頻度情報を元に画像に変換したり，パケットトラフィックデータを２次元平面上に従来の自己組織化マップを用いて写像した図に変換し,CGH-SOM でトラフィックデータの変化を解析する方法について研究を進める．

・ 構造的なモデルをマルチモーダル化した自己組織化マップの研究

これまでのマルチモーダルなベクトルを用いたシーケンスの学習に，隠れマルコフモデル等の構造的なモデルを組み合わせた入力を用いるパレート学習型 SOM を用いてシーケンスの学習を行い，その結果の評価を行い，DNA シーケンスの解析システムの開発を進める，また，株価情報等時系列データの解析に隠れマルコフモデルを用いた自己組織化マップを応用する．

・ 新たな応用分野の開拓

パレート学習型 SOM の拡張として、SOM，SP-SOM，P-SOM の適用可能な新たな応用を開拓する。現在では，セキュリティ関連において，大規模な WEB ページ解析，ウイルス，ワームなどのコード解析，セキュリティに関するログファイルの解析などを考えている．

４．研究成果
・DNA シーケンス解析

2012 年度に DNA シーケンス解析においては，まず，DNA コンテキストを用いた DNA シーケンス解析に関する研究を行った．DNA コンテキストはそれまで DNA 解析に用いられてきた頻度情報の拡張にあたるもので，モチーフ情報等の抽出が可能になり，より，解析精度が上がることが期待された．ただ，実際に実験を行ってみると，頻度情報とほぼ同じ程度の解析精度となった．

2013 年度は隠れマルコフモデルをユニットとする SOM(HMM-SOM)を用いた解析手法について研究を行った。ただ、HMM-SOM はシーケンスデータをそのまま学習させると、計算量が多くなり、大規模データ解析には適用できなくなるので、まず、SOM で一定長のプローブを生成し、プローブごとの頻度情報を用いて、HMM を生成する手法を開発し、実験を行った。結果としては頻度情報では分類不可能であった遺伝子グループごとの分類が可能であることがわかったが、やはり、HMM-SOM の計算量は大きく、HMM 自体の学習アルゴリズムに新たな手法が必要であると考えられる。

２０１４年度には塩基間の相関係数を用いた解析手法について研究を行った。相関係数を用いた方法は、従来の頻度を用いた手法と比較すると、長さ１０００以上のシーケンスデータに関しては、少ない次元数の入力ベクトルで同等かそれ以上の分類制度を持つことがわかったが、次世代シーケンサーの生データでの解析を考え、シーケンス長が短くなると、頻度情報より高速に精度が落ちることがわかった。また、多塩基配列の相関係数を用いることで、精度が向上するが、やはり、配列長が短くなると劣化の速度が速いことがわかった。また、マルチモーダル情報として、頻度情報やコンテキストと相関係数を組み合わせることで、ある程度分類精度を向上させることが可能であることが示された。これらの研究成果は国際会議の予稿集および Springer の Lecture note に掲載されている。

・IP パケット解析

2012 年度は IP パケットヘッダの各情報の頻度情報やペイロードデータの８ビットコードの頻度情報を用いて IP パケットトラフィックの可視化および異常パケットの検出の実験を行った。可視化については以前にも同様の実験を行っているが、異常パケットの検出に関してはより現実的な状況として、DOS-Attack と DDos-attack を想定して実験を行った。結果としては以前よりもは検出精度が落ちたが、ある程度の検出が可能で、より現実的な状況での実験としての結果が得られた。

2013 年度には前述の CGH-SOM を用いたトラフィックの可視化と異常検出の実験を行った。CGH-SOM においても SOM と同様にトラフィックの可視化が行え、また、異常検出も可能であった。また、Dos-Akkack に関しては SOM よりも大きく精度が改善されたが、DDos-attack に関しては精度がおちる結果となった。ただ、SOM の検出精度が逆の特性を示すため、２つを組み合わせて用いることで、ある程度実用的な検出が可能であると考えられる。なお、２０１４年度は担当者がいなかったため、本研究はペンディング状態であり、国内のニューラルネットの会議のみ発表を行った。

・　マルチモーダル生体認証システム

2012 年度にタッチポネルに書いた一筆書きの図形から、特徴点を自動的に検出し、特徴点間の筆速を生体情報として、特徴点を結合する順番を知識情報として用いるマルチモーダルな認証システムを提案し、現在までその応用について研究を進めている。ただ、以前の認証システムでは筆速の分布を生体情報としていたため、次元数が高くなり、高度な認証方式が必要となったため、SOM を用いていたが、今回の方式では特徴点間の速度を用いるため、高々７、８次元の生体情報であり、ユークリッド距離による判定で十分な精度が得られたため、SOM は適用していない。本研究はセキュリティの国際会議などで発表を行い、論文執筆のため実験結果を収集中である。

・　HMM-SOM を用いた時系列解析

・C 言語学習時の履修者の状況の可視化

新たな応用分野として C 言語の学習状況の可視化の実験を行った。履修者のキーボードとマウスの入力を一定間隔でサーバーに収集し、サーバー上でリアルタイムに自己組織化マップを用いて可視化し、履修者の状況を確認できるようなシステムの構築を目指している。現在のところ、実際に講義中に収集したデータの可視化をオフラインで行い、学習状況の可視化が行えることを確認した、今後、オンラインで使用できるシステムの構築を行う予定である。また、現在までの実験結果より約１０００名くらいのオンラインクラスまで適用可能であると考えられる。本研究は国際会議で発表している。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

①Gen Niina, Hiroshi Dozono: The Spherical Hiddeen Markov Self Organizing Map for Learning Time Series Data, Artificial Networks and Machine Learning, LNCS7562, vol2,563-570,2012、査読あり

② Gen Niina, Hiroshi Dozono, Kazuhiro Muramatsu, The Frequency Integrated Spherical Hidden Markov Self Organizing Map for Learning Time Series Data, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.19, 212-216,2015, 査読あり

③ Hiroshi Dozono: Visualization and Classification of DNA sequences using Pareto learning Self Organizing Maps based on Frequency and Correlation Coefficient, Advances in Self Organizing Maps and Learning Vector Quantization, Advances in Intelligent Systems and Computing, Vol295, 89-98, 2014, 査読あり

〔学会発表〕（計 5 件）

① Hiroshi Dozono, Takayuki Inoue, and Masanori Nakakuni: A Study of User Interfaces for Biometric Authentication System, The 2012 International Conference on Security and Management, July, 2012、Lasvegas, USA, 査読あり

② Hiroshi Dozono, Kentaro Kaneko: Visualization of Relations among DNA sequences using Self Organizing Maps, Eurpean Conference of COMPUTER SCIENCE, Dec. 2012、Pari, France, 査読あり

③Hiroshi Dozono, Gen Niina: Mapping of

DNA sequences using Hidden Markov Self Organizing Maps, IEEE symposium on Computational Inteligence, Apr. 2013, Singapore, 査読あり
④ Hiroshi Dozono, Masanori Nakakuni: Support System for C Programming Instruction in Group Education Environment, Modern Computer Applications in Science and Education, 14th International Conference on Computer Supported Education, Jan.2014, Boston, USA, 査読あり
⑤ Hiroshi Dozono:Visualization of the Sets of DNA sequences using Self Organizing Maps based on Correllation Coefficient, 2013 IEEE International Conference on Bioinformatics and Biomedicine, Dec. 2013, Shanghai, China, 査読あり
〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕
ホームページ等
http://www.fusion.saga-u.ac.jp/research/douzono.html
６．研究組織
(1)研究代表者
堂薗　浩（Hiroshi Dozono）
佐賀大学・工学系研究科・准教授
研究者番号：00217613

(2)研究分担者
中國真教（Masanori Nakakuni）
福岡大学・
総合情報処理センター研究開発室・准教授
研究者番号：10347049

(3)連携研究者　なし

(4)研究協力者
新名　玄　佐賀大学工学系研究科博士後期課程３年

岡田望邦　佐賀大学工学系研究科先端融合工学専攻平成２５年度修了