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　５ 年　 ６ 月 １４ 日現在

機関番号：１７６０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ１２１３９

研究課題名（和文）数理モデルと機械学習を組み合わせたスモールデータ処理基盤技術の構築

研究課題名（英文）Development of small data processing method combined with mathmatical model and machine learning approarch

研究代表者
山森　一人（Yamamori, Kunihito）

宮崎大学・工学部・教授

研究者番号：５０２９３３９５

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：現在の機械学習では，学習に膨大なデータを要する．本研究では，観測や再現が困難な問題について，現象を近似する数理モデルと機械学習を組み合わせて解決するアプローチを試みた．題材としては，研究代表者が共同研究者として関わった，たんぱく質発現量から生理活性値を推定する問題，およびデータ数は多いもののその信頼性に疑義がありデータ数を削減せざるを得ないネットワークセキュリティ分野における侵入検知問題を扱った．前者については仮想データ生成までは行ったものの，推定プログラムは開発途中である．後者について，データ数を縮減しつつ学習結果の精度評価を行い，GBDT系アルゴリズムが優秀であることを示した．

研究成果の学術的意義や社会的意義
現代のAIでは，適切な結果を得るためには膨大な数の学習サンプルを必要とする．一方，観測が困難であったり再現が難しいなど，多数の学習サンプルを準備することが難しい課題も存在する．本研究では，学習データを補うため数理モデルを作成し，モデルに従って学習サンプルを生成することで精度よく学習が行うことが可能なアプローチを模索した．例題として，たんぱく質発現量から生理活性値を推定する問題，および学習データ数は豊富なものの信頼性に疑義があるコンピュータシステムへの侵入検知問題を取り上げた．前者については推定プログラムを作成し，後者についてはブースティングを併用した決定木アプローチが有効であることを示した．

研究成果の概要（英文）：Recent machine learning algorithms require a large number of training samples. In this research, I try to combine the mathematical model that can approximate the phenomena and the machine learning approach to solve some problems that are hard to observe the phenomena or hard to reproduce the experimental results again. I picked up two problems; one was to estimate the physiological activities from the protein expression levels, and the other was to detect the intrusion into the computer systems. For the first one, I rewrote the Linux-based programs into an integrated program. For the second one, I showed Gradient Boosted Decision Tree algorithm was suitable and robust for the small number of training samples.

研究分野：ソフトコンピューティング

キーワード：機械学習　スモールデータ　数理モデル　決定木　ネットワークセキュリティ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

ビッグデータを背景に，機械学習（AI）の様々な分野への応用が試みられている．調節可能な多数の結合荷重を内包する多層神経回路網では，その結合荷重数に応じた，適切かつ大量の学習サンプルを準備する必要がある．例えば，世界最大の画像分類ベンチマークデータセットである Google Open Image V4 では 170 万枚を越える画像において，600 種類のオブジェクトに対しラベルを付与している．神経回路網を用いたパターン分類やパターン認識では，神経回路網の規模に比べ学習サンプル数が少ないと学習サンプルに過度に適応した学習が行われてしまい，未知入力に対して望ましい出力が得られない，過学習と呼ばれる問題が発生する．つまり，現代的機械学習アプローチには，多数の適切な学習サンプルの収集がその第一ステップとなる．

一方，十分な数の学習サンプルが収集できない分野も存在する．研究代表者はかつて農学系の研究者と組み，たんぱく質発現量からの食品機能性推定に関する研究に従事した．この研究では，まず培養がん細胞に薬品等を作用させた後，細胞内のたんぱく質発現量 13 種類と，培養がん細胞の増減（機能性）をそれぞれ測定し，これらの間の入出力関係（入力：たんぱく質発言量，出力：がん細胞の増減）を学習により獲得する．その後，作用が未知の食品抽出物（煮汁やアルコール抽出液）をがん細胞に作用させてたんぱく質発現量のみを測定，学習済みの推定器によりその機能性を数値として推定するものである．培養がん細胞の増減は研究員が顕微鏡を覗きつつ目視でカウントしており，多数回の測定を実施できる環境ではない．濃度の高い薬品等を作用させたときは一部の細胞が死滅してしまい，他とは大きく異なる測定値が観測されることもあった．さらに，たんぱく質発現量と機能性を測定する培養がん細胞は複数の異なるシャーレ上に分配されており，それぞれのシャーレから測定値が得られるものの，複数の測定値は互いに対応付けられてはいない．いわば，身長と体重の相関図を描くのに，それらを関連付ける氏名や番号が欠けた状態であった．

これらの経験をまとめると，以下の問いに収斂する．

- 比較的大きなノイズ（測定誤差等）を含むデータから，信頼できるデータをどのように選び出すのか？
- 未整理，かつ少数の測定値をいかに有効に利用し，機械学習のアプローチにのせていくのか？
- 少ない学習サンプルでも過学習を起こさず，未知データに対しても信頼性の高い解が得られる学習法や学習モデルはどのようなものか？

研究代表者は，検定による外れ値の除外，回帰分析による入出力の対応付け，拡張重み更新型自己組織化マップを考案しての学習と推定という 3 つの技術により，この課題に対応した．本研究課題では，このアプローチを発展させ，より精度を高める技術を開発すること，および分野の研究課題にも応用範囲を拡げることを目的とした．

２．研究の目的

本研究の目的は，未整理かつ少数のデータしか得られていない環境において，現代的機械学習アプローチで必要となる「適切な」学習サンプルを構築する手法，およびそれらを適切に学習し汎化できる学習モデルを構築することとした．

３．研究の方法

学習データが少ない課題としては，従来から取り組んでいたたんぱく質発現量からの生理活性推定に用いていたデータを活用し，数理モデルを取り入れた仮想データ生成システムを構築することとした．これにより仮想的なデータを生成して学習を行うことで，推定精度の向上を目指した．さらに，複数のプログラム群から構成していた推定システムを統合した，一貫したシステムとして再構成することとした．加えて，入力-結合荷重間の距離関数として一般的な，ユークリッド距離以外を用いることで精度向上が可能か検証することとした．

別の課題への応用として，データ数は多いもののラベルが不正確である可能性のある，コンピュータシステムへの侵入検知を例題とし，ラベル付けの再検討，学習サンプル数を縮減しての検出精度の検証，各種機械学習アルゴリズム間の精度比較を行うこととした．

４．研究成果

研究期間中，一時体調を崩し半年ほど研究を休止せざるを得ない期間があったものの，コンピュータシステムへの侵入検知に対する学習データ縮減の影響については調査を終え，途中経過も含め国際会議にて発表を行った．データ縮減については，GBDT（Gradient Boosted Decision Tree）系のアルゴルムが比較的ロバストであること，従来用いられてきた侵入検知学習用データセットのラベルについて，扱いには注意が必要であることを示した．生理活性推定については，複数のプログラムを組み合わせていた従来から，統合したプログラムへの刷新を行ったが，これ

についてはデバッグが完了しておらず，今後も継続して開発を行っていく．

上記の理由から，本研究成果報告ではデータ縮減に対するロバスト性について記述する．従来，この分野では KDD CUP99 と，これを元に作成された NSL-KDD というデータセットが用いられてきた．このデータセットはコンピュータシステムへの攻撃を模した通信で作成されたものであり，現実世界での攻撃通信を忠実に反映しているとは言えない．そこで，京都大学に設置されたハニーポットに対する通信を収集し，新たに Kyoto 2006 データセットが作成され，これにラベル付けの修正，およびデータ収集期間を拡張した Kyoto 2016 データセット（以下，K2D と略す）を本研究の題材に選択した．

はじめに，K2D と今後のデータ解析に便利なように RDBMS（Relational Database Management System）上に格納した．K2D の元論文では，データ収集期間を 2 ヶ月毎に区切り，決定木やランダムフォレストなどのアルゴリズムにより正解率，適合率，検知率，誤検知率を評価している．本研究課題では，K2D を NIDS に用いるにあたり，以下の観点からデータを期間で区切ることをせず，全体を通じて統計的な検知から解析を行った．

- コンピュータシステムへの攻撃は，時間経過により手段が移り変わるものの，そのタイミングは予見できず，2 ヶ月という期間が妥当な否か十分な検討が行われていないこと
- 機械学習アルゴリズムの「学習」という観点から，K2D は多数のサンプルを格納しているものの，ラベル付けの正確さやプロトコル間のデータ数の不均衡などについても十分な検討が行われていないと考えられること

４．１．プロトコル間の不均衡

K2D には，TCP，UDP，ICMP の 3 種類のプロトコル通信が含まれており，それぞれ「既知攻撃」,「未知攻撃」,「正常通信」の 3 種類のラベルが付与されている．図 1 は各ラベルに含まれる

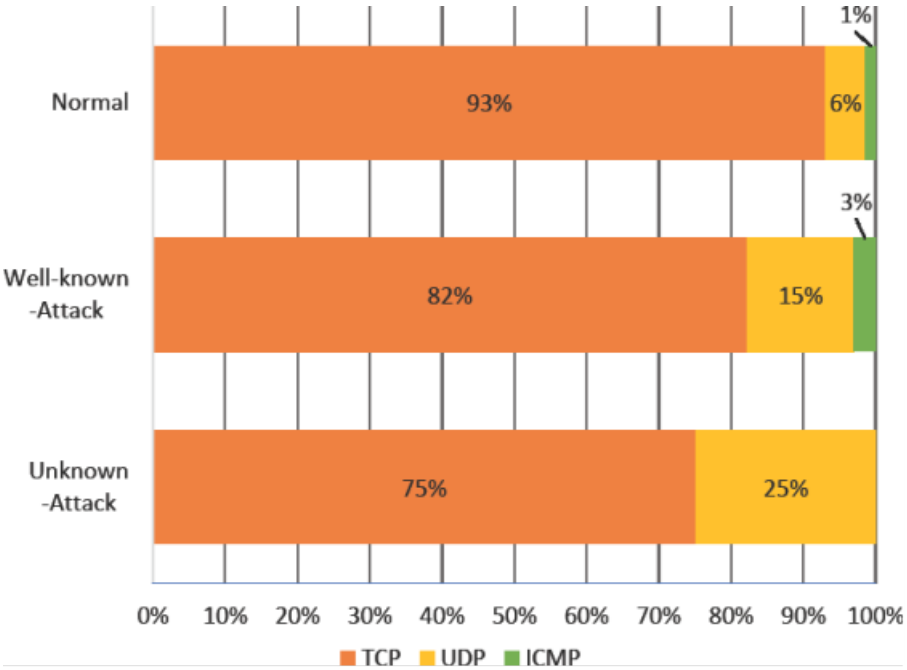

図 1　ラベル毎のプロトコルの割合

プロトコルの割合を示したものである．図 1 から分かるように，各ラベルで TCP がその多くを占め，ICMP はほとんど含まれていない．このことから，単純にデータセットからデータをランダムに抽出し学習用データを構成すると，TCP に偏った学習を行ってしまう恐れがある．そこで，本研究課題では K2D 中で多数を占める TCP を対象に解析することとした．

４．２．データの重複

次に，同一プロトコル，同一ラベルのデータにおいて，データベースを精査したところ，まったく同一のデータが多く含まれていることが分かった．これは DoS（Denial of Service）攻撃など，短時間で多くの攻撃パケット（セッション）がコンピュータシステム上に到達しているような場合，同一の特徴量を持つデータが多く観測されるようなケースと考えられる．一方，機械学習における学習用データと考えた場合，同一のデータが多く含まれていることは，実質的な学習サンプルの種類が減少してしまうことを意味する．そこで本研究課題では，以降のデータの解析と学習用データセットの構築においては，同一ラベルにおいて同一の特徴量を持つデータは予め取り除くこととした．

４．３．同一特徴量異ラベルデータ

重複を取り除いたデータセットにおいて，ラベル間のデータを比較した．その結果，まった

く同じ特徴量を持つデータが異なる分類にラベル付けされているケースが発見された．K2D のデータが実際の通信に基づいて作成されており，パケットそのものではなく，パケット（セッション）から得られる特徴量をサンプルとしていることからあり得ることではあるが，機械学習の学習用データとするには不適切である．そこで，K2D をベースに機械学習用データセットを再構成するにあたり，これらの同一特徴量を持ちつつ異なるラベルが付与されているデータについては，どちらのラベルが正解であるか不明でもあり，一律に取り除くこととした．

４．４．特徴量とラベルの不一致

K2D のラベルが「既知攻撃」,「未知攻撃」,「正常通信」の 3 つであることは先述した．このうち「既知攻撃」,「未知攻撃」のラベルはハニーポットへの通信に，「正常通信」のラベルはハニーポット以外に準備したホストへの通信に与えられている．さらに「既知攻撃」と「未知攻撃」は，特徴量に含まれる 3 つの攻撃検知ツールのうち，特定のツールのみが「攻撃」と検知したセッションを「未知攻撃」と分類している．すなわち，特徴量のなかに「攻撃検知ツール」による「検知結果」が含まれている．そこで，通信を受けたホストに基づいて決定されたラベルと，特徴量に含まれる攻撃検知ツールの検知結果を突き合わせ，ラベルの妥当性について検討した．この結果を表 1 にまとめる．

表 1　ラベルと攻撃検知ツールによる検知結果との比較

ラベル	検知ツール検出結果		計
	検知なし	検知あり	
正常通信	142, 164, 176	7, 567, 174	149, 731, 350
既知攻撃	516, 989, 852	9, 392, 663	526, 382, 515
計	659, 154, 028	16, 959, 837	676, 113, 865

表 1 に示したように，TCP における「正常通信」ラベルデータ，約 1 億 5 千万のうち，およそ 750 万のデータが何からのツールにより攻撃的通信と判定されており，「既知攻撃」ラベルデータ約 5 億 2 千 6 百万のうち 5 億 1 千 6 百万のデータはどの攻撃検知ツールによっても攻撃的通信と判定されていないことが分かった．これらのことから，K2D 全体をそのまま機械学習における学習データとするには問題があるという結果が得られた．

４．５．少数サンプルに対する機械学習アルゴリズムの頑健性評価

重複したサンプルやラベルに矛盾があるデータを除いたうえ，複数の機械学習アルゴリズムにおいてサンプル数減少に対する頑健性を調査した．なお，K2D を用いて NIDS を構成し機械学習アルゴリズムを比較評価した既存研究はそれほど多くないことから，従来から本研究分野で多く用いられてきている NSL-KDD データセットを用いた．

比較対象としたアルゴリズムは GBDT (Gradient Boosting Decision Tree) 系から XGBoost,

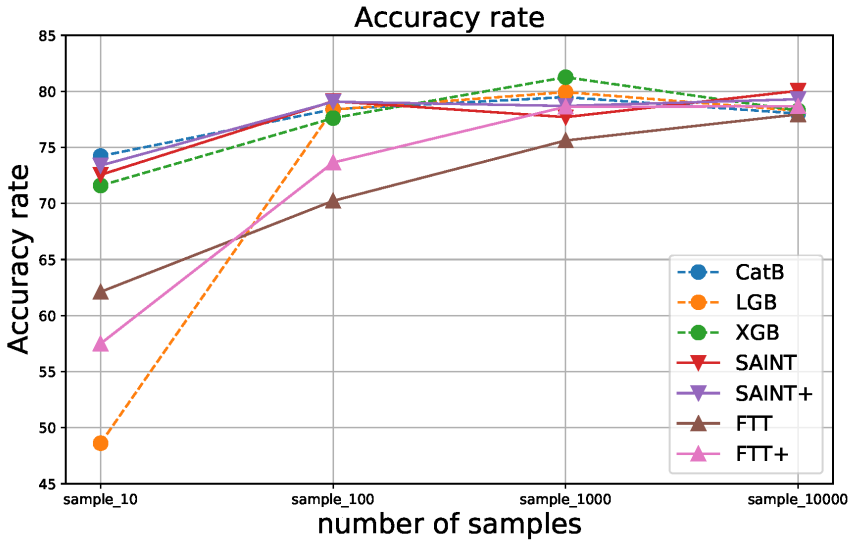

図 2　学習サンプル減少に対する機械学習アルゴリズムの正解率の推移

LightGBM, CatBoost の 3 種類, 深層学習系から SAINT (Self Attention and Intersample Attention Transformer), FT-Transformer の 2 種類に，PLE (Piecewise Linear Encoding) の有無を加えて 4 種類の計 7 種とした．なお，サンプルが十分準備できる場合，これらのアルゴリズムの検出

結果にそれほど大きな差はない.

図 2 に，サンプル数を 10 から 1,000 まで変化させたときの，各アルゴリズムでの正解率を示す．図 2 から，学習サンプル数が 10,000 を越えれば，どのアルゴリズムも正解率は 80%程度となり差はないものの，サンプル数減少に対する頑健性については，GBDT 系アルゴリズム（CatBoost，XGBoost）が比較的良い結果を示すことが分かった.

５．まとめ

本研究課題では，現在盛んに研究されており一部実用化も図られている深層学習アルゴリズムを適用できない，学習サンプル過少，あるいはサンプル取得が困難な課題に対し，どのような学習アルゴリズムが適しているのか，少ない学習サンプルを如何に活用し妥当な学習結果を得るかについて研究を行った.

対象とした課題としては，たんぱく質発現量からの食品機能性推定，および NIDS 用データセットを用いてのサンプル数削減を取り上げた．研究代表者が一時期体調を崩してしまい，前者の課題についてはプログラムのデバッグ途中という結果となったが，後者の課題については既存のデータセットの問題について指摘を行ったこと，および既存データセットにおいてサンプル数を変化させて 7 種類の機械学習アルゴリズムの頑健性について評価を行った.

今後の課題としては，前者のデバッグを完了させ一貫したシステムとして完成させること，および NIDS に適したデータセットの整備と学習アルゴリズムの開発が挙げられる.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　1件／うちオープンアクセス　0件）

１．著者名 Ryo SAITO, Masaru AIKAWA, Kentaro INOUE, Kunihito YAMAMORI	４．巻 25
２．論文標題 Affect of data unbalance in "Kyoto 2016 Dataset" for NIDS with machine learning	５．発行年 2020年
３．雑誌名 Proc. International Symposium on Artificial Life and Robotics	６．最初と最後の頁 612-616
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 該当する

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　5件）

１．発表者名 Kiyoko Nagahama, Akira Ota, Katsuhisa Kurogi, Kunihito Yamamori, Yoichi Sakakibara
２．発表標題 Simultaneous estimation of multiple food functions of food components using a targeted proteomics approach
３．学会等名 10th Asia-oceania Human Proteome Organization Congress（国際学会）
４．発表年 2021年

１．発表者名 Akinobu Iwai, Masaru Aikawa, Kunihito Yamamori
２．発表標題 Driving trajectory optimization by reinforcement learning for motorsports
３．学会等名 The 27th International Symposium on Artificial Life and Robotics 2022（国際学会）
４．発表年 2022年

１．発表者名 Kosuke Yoshida, Masaru Aikawa, Kunihito Yamamori
２．発表標題 Heuristic base music arrangement suppressing on discord progression
３．学会等名 The 27th International Symposium on Artificial Life and Robotics 2022（国際学会）
４．発表年 2022年

１．発表者名
永濱清子，太田輝，黒木勝久，山森一人，水光正仁，榊原陽一

２．発表標題
ターゲットプロテオミクスによる複数の食品機能性の同時推定

３．学会等名
第27回　日本生物工学会九州支部　大分大会

４．発表年
2021年

１．発表者名
Ryo Saito, Kunihito Yamamori, Masaru Aikawa, Kentaro Inoue,

２．発表標題
Network Design for Session Type NIDS

３．学会等名
The 26th International Symposium on Artificial Life and Robotics 2021（国際学会）

４．発表年
2021年

１．発表者名
Chihiro Kudo, Kunihito Yamamori, Masaru Aikawa, Kentaro Inoue, Ryo Saito

２．発表標題
Tuning Support Tool for WAF Mod Security by Log Analysis with Machine Learning

３．学会等名
The 26th International Symposium on Artificial Life and Robotics 2021（国際学会）

４．発表年
2021年

１．発表者名
齋藤燎，相川勝，井上健太郎，山森一人

２．発表標題
「Kyoto 2016 Dataset」における冗長性と同一特徴量異ラベルデータに関する報告

３．学会等名
2019年度（第72回）電気・情報関係学会九州支部連合大会

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