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　６ 年　 ６ 月 ２０ 日現在

機関番号：２２７０１
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１２０３２

研究課題名（和文）オンライン予測理論に基づくデータサンプリング問題への統合的アプローチ

研究課題名（英文）Unified approach for data sampling problems based on online prediction theory

研究代表者
末廣　大貴（Daiki, Suehiro）

横浜市立大学・データサイエンス学部・准教授

研究者番号：２０７８６９６７

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：機械学習における様々なデータサンプリング問題に対し，オンライン予測理論に基づく統合的定式化と理論解析を行った．具体的には，Lerning from Label Proportions と呼ばれる学習問題における疑似ラベル選択，ノイズラベルあり学習問題におけるノイズデータ回避を考え，学習器の挙動に応じて適応的にデータをサンプリングする統合的な枠組みを構築した．いずれの問題においても理論的に適切なサンプリングが行えることを証明し，かつ実験的にも最新手法を超える性能を達成することを示した．

研究成果の学術的意義や社会的意義
データから学習を行う機械学習は人工知能の中核をなす技術である．一般に，データに付与される「正解」は誤り（ノイズ）が含まれていたり，全てのデータに付与されていなかったり，不完全なものであることが多い．このようなデータから適切な学習を行うためには，データ集合の中から適切な情報だけを取り出すサンプリングが重要な役割を担う．しかし，サンプリングはデータの性質やタスクに応じたアドホックな定式化や手法が多く，汎用性や理論解析に関する議論が欠如していた．本研究ではデータやタスク依存の現状を打破する統合的な枠組みと理論性能保証の指針を与え，サンプリング技術ひいては機械学習技術の発展に大きく寄与するものである．

研究成果の概要（英文）：For various data sampling problem in machine learning, I designed a unified formulation and gave theoretical analyses based on online prediction theory. More precisely, for the pseudo labeling problem in Learning from Label proportions and data selection problem in learning with noisy labels, I proposed a unified framework for adaptively sampling good data according to the learning behavior. For both problems, I proved the proposed algorithms work effectively in theory and in practice.

研究分野：統計的学習理論，オンライン意思決定理論，機械学習応用

キーワード：データサンプリング　オンライン予測　ノイズあり機械学習

1. 研究開始当初の背景

機械学習に用いられる膨大なサンプルの中には，**学習器にとって望ましいデータ，望ましくないデータ，どちらも含まれている**．例えば，「教師ラベルが間違っているデータ（ノイズ）」や，「画質等が劣化しているデータ」のように，入力してしまうとモデルの汎化性能を下げてしまうデータや，予測基準のボーダー付近にいるような難しいデータ等，様々な性質のデータを含んでいる（**図１左**参照）．しかし，**学習器にとって望ましくないデータ，望ましいデータを事前にフィルタリングすることは簡単ではなく**，様々な方法が提案されている．

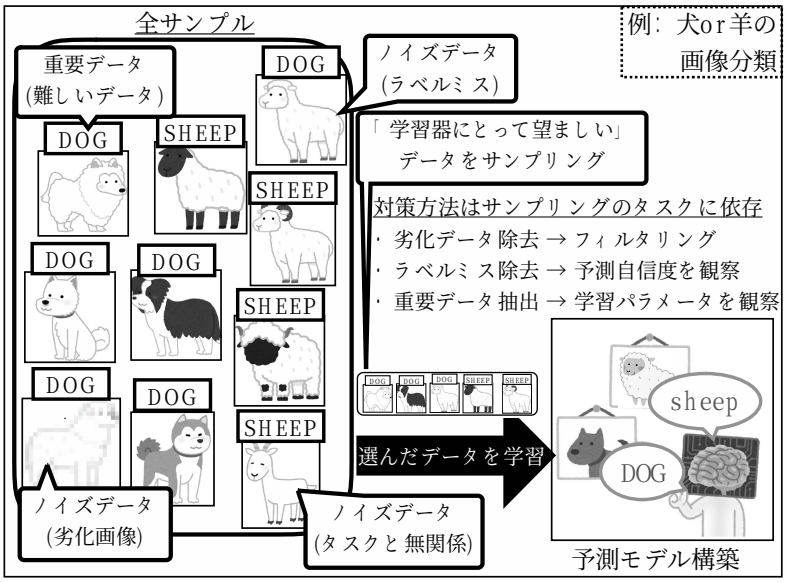

図 1: 画像データに対するデータサンプリングの例

① 「良いデータサンプリング」とは？ ～統合的な定式化～

データサンプリングには，ドメイン（医療画像，センサデータなど）やタスク（ノイズ除去，重要データ抽出など）に応じて様々なものが存在する．様々なドメインで共通のサンプリングタスクであり，これまでに最も多く検討されているのが，「ノイズデータの除去」というタスクである．しかし，ノイズと言っても一概ではなく，一部の教師ラベルが間違っていると言ったケースや，タスクに関係のないデータが混在しているケース，質の低いデータが混在しているケースなど，様々なものがある（**図 1 左**参照）．既存研究では，ドメイン，タスクの細かい特性に応じて，様々な定式化や手法の提案がなされている（**図１右**参照）．

しかし，**ドメイン・タスク依存の手法はアドホックなアプローチが多く**，データサンプリングの一般定式化や，汎用性に関する議論はなされていない．データサンプリング技術は，様々な分野で必要な技術であるが，汎用性の欠如は技術の普及に大きな障害となる．そこで，**ドメイン・タスクによらない統合的な枠組み（定式化および手法）を開発することで汎用性の向上と理論解析を促進**する必要がある．

②高精度かつ効率的なサンプリング方法は存在するか？～理論性能保証～

フィルタリングや異常検知などの古典的なものから，学習器の内部パラメータ情報を観察した方法まで様々なデータサンプリング方法が提案されているが， 理論的な見地は与えられていないものが多い．理論解析は**人工知能技術の透明性，安定性を示す**上で非常に重要である．**サンプリングの精度及び計算量に関する性能を「ドメインやタスクに依存せず」理論的に保証**することができれば，人工知能技術の信頼性を大幅に上昇させることができる．

③実アプリケーションの範囲は？～実応用の開拓～

統合的な定式化を与えることができれば，汎用性が上がり，アプリケーションを拡大することができると考えられる．従来データサンプリングが行われてきたドメインや，従来考えられていたデータサンプリングタスクにとどまらず，**新しい実アプリケーションを開拓し，アルゴリズムが実用的な精度，計算時間で動作する**を示すことは，非

常に重要である.

2. 研究の目的

データサンプリング問題とは, ユーザが用意したある学習器に対し, どのようなデータをサンプリングすれば学習器が良いモデルを構築できるか, を考える問題である. 本研究では, データサンプリング問題に対するドメイン, タスク依存の現状を打破し, **①データサンプリング問題の統合的枠組みの構築**, **②精度および計算量に関する理論性能を保証したアルゴリズムの提示**, **③実応用の開拓**を目的とする.

3. 研究の方法

本研究の根幹は「**データサンプリング問題の最も素朴な解き方**は, 学習データの全部分集合(サンプリングの全候補)を考え, 対応する予測モデルを全て構築し, 比較評価することである」という点にある(**図 3**). しかし, 全部分集合はサンプル数を n とすると 2^{n-1} 通り存在し, それら全てに対し予測モデルを構築することは**計算量的に現実的でない**.

そこで本研究では, **データサンプリング問題を, オンライン組み合わせ集合予測問題として捉える**. 具体的には, 組み合わせ予測アルゴリズムが, 「望ましい組み合わせ集合(＝望ましいデータ集合)」をオンライン(逐次的に)選択し学習機に与え, 学習されたモデルの振る舞いを評価しながら組み合わせ予測ルール(サンプリングルール)を更新することで「望ましいデータ」と, それに基づくモデルを構築していく(**図 4**).

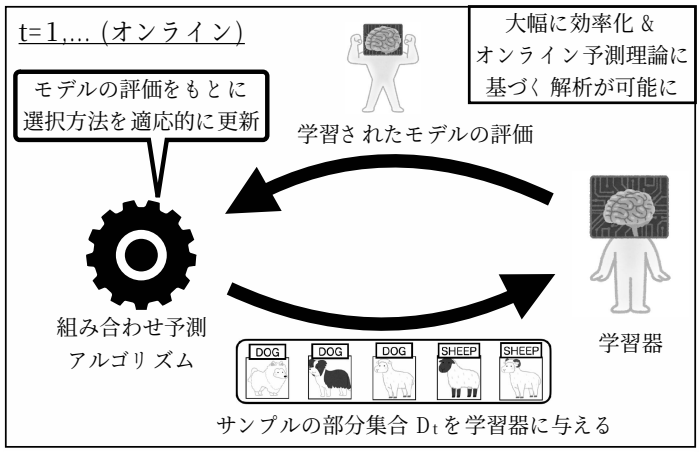

図 3: オンライン組み合わせ予測問題としての枠組み

②オンライン予測理論に基づく理論性能の保証

①のアプローチにより, **「オンライン予測理論」に基づく理論解析**が可能になる. オンライン予測理論では, 大まかに言うと**図 4** のような繰り返し予測問題に対し, 「リグレット(後悔)」と呼ばれる指標を小さくすることを目的とする. 本研究においては, 十分に繰り返しを行ったあと「この組み合わせ集合を選んでおけばよかった」という後悔に対応する. すなわち, リグレットを小さくすることで, **最適な組み合わせ集合に匹敵するようなデータサンプリングが実現できる**. また, 申請者の過去の研究[研究実績 14]などから, このような**組み合わせ構造に対しても効率的に予測可能なアルゴリズムが存在**し, 活用が見込める.

② 実応用の開拓

様々なドメイン・タスクに適用しながら, ②で示した理論性能を保つことができるか, 評価を行う. まずは, 従来最も研究が行われているノイズデータ回避への応用を行い, 実用性を証明する. **ドメイン知識の複雑性が高い, 医療データやスポーツに関するデータなどへの応用**を検討し, ドメインの拡大を図る. また, **新たなデータサンプリングタスク**として, データの学習難易度を目標の性能に応じて自動サンプリングすることによる学習器の性

能制御，分類問題における，重要クラスデータの自動重点サンプリングなど，**ノイズ除去に限らない様々なタスクを創出，解決可能**であると考えている．

4．研究成果

機械学習における様々なデータサンプリング問題に対し，オンライン予測理論に基づく統合的定式化と理論解析を行った．具体的には，分類学習における性能制御問題，Lerning from Label Proportions（LLP）と呼ばれる学習問題における疑似ラベル選択，ノイズラベルあり学習問題におけるノイズデータ回避を考え，学習器の挙動に応じて適応的にデータをサンプリングする統合的な枠組みを構築した．いずれの問題においても理論的に適切なサンプリングが行えることを証明し，かつ実験的にも最新手法を超える性能を達成することを示した．

性能制御問題に関しては，画像認識･パターン認識に関する国内最大の会議 MIRU2022 にオーラル発表として採択され，発表を行った．LLP に関する研究成果は，信号処理分野におけるトップ会議である ICASSP2023 で論文および成果発表を行った．また，ノイズラベルあり学習問題における研究成果は機械学習分野で権威のあるジャーナルである Machine Learning に論文が掲載され，ECML/PKDD2023 で発表も行った．

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Song Heon、Mitsuo Nariaki、Uchida Seiichi、Suehiro Daiki	４．巻 113
２．論文標題 No regret sample selection with noisy labels	５．発行年 2024年
３．雑誌名 Machine Learning	６．最初と最後の頁 1163～1188
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10994-023-06478-8	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 Matsuo Shinnosuke、Bise Ryoma、Uchida Seiichi、Suehiro Daiki	４．巻 -
２．論文標題 Learning From Label Proportion with Online Pseudo-Label Decision by Regret Minimization	５．発行年 2023年
３．雑誌名 Proceedings of the 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing	６．最初と最後の頁 1～5
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICASSP49357.2023.10097069	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 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 本田康祐，内田誠一，末廣大貴
２．発表標題 識別器の斟酌学習
３．学会等名 電子情報通信学会 パターン認識・メディア理解研究会（PRMU研究会）
４．発表年 2021年

１．発表者名 本田 康祐，内田 誠一，末廣 大貴
２．発表標題 識別器の斟酌学習
３．学会等名 画像の認識・理解シンポジウム(MIRU2022)
４．発表年 2022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