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 ２０ 日現在

機関番号：１４３０１
研究種目：挑戦的研究（萌芽）
研究期間：2017～2020
課題番号：１７Ｋ１９９７３

研究課題名（和文）機械学習アルゴリズムのための離散データ上の関数に対する解析的最適化数理の構成

研究課題名（英文）Construction of mathematical optimization methods for discrete data useful in machine learning algorithms.

研究代表者
山本　章博（Yamamoto, Akihiro）

京都大学・情報学研究科・教授

研究者番号：３０２３０５３５

交付決定額（研究期間全体）：（直接経費）　　4,900,000円

研究成果の概要（和文）：機械学習は自然言語データの処理における基本技術となっている．自然言語データを数ベクトルのデータに変換した上で，深層学習など最新の機械学習技術を適用する方法は学習結果の意味を解釈しづらく，さらには文のもつ自然な構造がベクトルという平坦な構造で適切に表現できる保証はない．本研究では，自然言語データの数理的なモデルである文脈自由言語の構文解析木，一階述語論理の文，語の列の直接の代数化であるパターンを対象とした機械学習のための最適化数理とアルゴリズムを構築した．

研究成果の学術的意義や社会的意義
機械学習は自然言語データの処理における基本技術となっている．特に自然言語データを自然数ベクトルのデータに変換した上で，深層学習など最新の機械学習技術を適用する方法は大きな成果を上げつつある．しかし，深層学習は学習結果の意味を解釈しづらく，さらには文のもつ自然な構造がベクトルという平坦な構造で適切に表現できる保証はない．本研究で扱った，語の列である自然言語データ,あるいはそこから抽出した構文木を直接扱う機械学習アルゴリズムを用いれば，解釈可能な構造を表現した結果を出力することが期待される．

研究成果の概要（英文）：Machine learning is now a fundamental technology in processing data in natural languages. If we convert natural language sentences converted into vectors of number and then applied the latest machine learning techniques, such as deep learning, we would meet difficulty in interpreting the meaning of the learning results. Moreover, we would have no guarantee that the natural structure of a sentence are adequately represented with vectors whose structure is very flat. In this study, we have developed optimization mathematics and algorithms for machine learning for parse trees in context-free languages, which are mathematical models of natural language data, sentences in first-order predicate logic, and patterns, which are direct algebraic representations of word sequences.

研究分野：知能情報学

キーワード：機械学習　文脈自由文法　木構造　一階述語論理　帰納論理プログラミング　最小汎化

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

機械学習は自然言語データの処理における基本技術となっている．特に自然言語データを自然数ベクトルのデータに変換した上で，深層学習など最新の機械学習技術を適用する方法は大きな成果を上げつつある．しかし，深層学習は学習結果の意味を解釈しづらく，さらには文のもつ自然な構造がベクトルという平坦な構造で適切に表現できる保証はない．そこで，語の列である自然言語データあるいはそこから抽出した構文木を直接扱う機械学習アルゴリズムを構成すれば，解釈可能で文の構造を表現した学習結果が期待される．近年の機械学習理論では，“学習”を弁別関数とデータを引数とする関数の和の最小化問題として定式化する．本研究では，語列データあるいは構文木データである場合の機械学習アルゴリズムに利用可能な最適化理論を構成することを目標とする．

自然言語データあるいは自然言語データから抽出した係り受け構造を直接扱う機械学習の数理的基盤の確立を目標として，自然言語データに対する機械学習では，語彙や文脈を属性と見なした上で深層学習などの属性ベースの学習アルゴリズムを適用する方法がよく行われており，成果も多い．しかしながら，属性ベースの機械学習の結果として得られる弁別関数が何を意味しているのかを理解するのは困難である．深層学習に対しては，概念階層と脳構造の対比を主張しているが，それは弁別関数の意味を表しているわけではない．

本研究の代表者と分担者はこれまで確率文脈自由文法や確率 Horn 論理式の学習理論を構築してきた．それは，確率文脈自由文法が句構造文法の一種であり，確率 Horn 論理式は一階述語論理の論理式の一種であるであるため，意味構造を数理的に表現でき，人間にとっても解釈可能で，より高度な応用の可能性があるからである．しかしながら，機械学習を最適化問題と定式化すると，文脈自由文法や Horn 論理式の機械学習アルゴリズムを構築するためには，語列データや構文木データ上の関数の最適化理論を同時に構築しなければならなかった．属性ベースの機械学習アルゴリズムでは，実数値ベクトルに対する最適化アルゴリズムや離散凸解析によって構成された自然数値ベクトルに対する最適化アルゴリズムが利用される．後者は，離散データを対象とはするものの，文脈自由文法や Horn 論理式の機械学習に適用するには，語列データや構文木データを自然数ベクトルに変換する必要があり，その適切性を明示することが困難であった．その結果，語列データや構文木データ上の関数を直接扱う最適化理論が必要であるという認識に至った．

計算機科学としては，テキスト文書の構文木データなど構造的データに対して機械学習を行うアルゴリズムの基本部分を設計することになる．現在広く用いられている属性をベースにした自然数値ベクトルや実数値ベクトルからなるデータに対して行う機械学習とは異なる機械学習を行うアルゴリズムであって，応用範囲も学習の結果も異なり，新たな学問の展開が期待される．数理的側面としては，現在までに，自然数ベクトルに対して構築されてきた最適化理論である離散凸解析とは異なる対象となる離散データに対して新たな理論を構築するものである．

２．研究の目的

本研究の目的は，自然言語・形式言語の単語列・文字列(以下では語列という)あるいは構文木からなるデータの集合 D を対象とした機械学習において，弁別関数 f として句構造文法や論理式を採用した場合の機械学習に必要な最適化の数理的理論を構成することにある．現在，離散データを対象とした最適化理論が直接適用可能なのは D や f が自然数で表現される場合であり，それは全順序など実数から継承した自然数の性質を利用している．一方で，語列全体の集合，構文木全体の集合，句構造文法の集合，論理式の集合は，理論上は自然数に埋め込めるため最適化理論を適用化の杖はあるが，それらは実数や自然数とは異なる代数的・位相的構造を持つため，学習結果に意味解釈を与えるという動機を達成しない．本研究は新たな数理最適化理論の体系を構築に挑むことになる．

３．研究の方法

計算としての“学習”を弁別関数 f とデータ D を引数とする関数 $L(f, D)= \sum d(f, d)+ p(f)$の最小化問題として定式化しておく．ここで $\sum d(f, d)$は，f と D 中の各データ点 d との距離の総和であり，p(f)は関数の適切さを表す尺度である．D が語列データあるいは構文木データである場合の機械学習アルゴリズムに対しては，そのようなデータを対象とした最適化理論を構成して数理的基盤とする必要がある．本研究では，自然数ベクトル以外の離散データ構造上の最適化理論を構築することが目標である．特に，自然言語データと構文木データからの機械学習に利用しやすい最適化アルゴリズムを構築する基盤を構成する．

参考文献

[1] M. Nishino, A. Yamamoto, M. Nagata: A Sparse Parameter Learning Method for Probabilistic Logic Programs. AAAI Workshop: Statistical Relational Artificial Intelligence 2014.

[2] T. Yamazaki, A. Yamamoto, T. Kuboyama: Tree PCA for Extracting Dominant Substructures from Labeled Rooted Trees. Discovery Science 2015: 316-323.

[3] 山口，吉仲，山本:確率論理プログラムを用いた確率文法のパラメータ推定，第 93 回人工知能基本問題研究会，SIG-FPAI-98-07, 81-86, 2016.

[4] M. Nishino, J. Suzuki, M. Nagata: Phrase Table Pruning via Submodular Function Maximization. ACL (2) 2016.
[5] M. Nishino, N. Yasuda, S. Minato, M. Nagata: Zero-Suppressed Sentential Decision Diagrams. AAAI 2016: 1058-1066.

４．研究成果

[1] 文脈自由言文法に従う語列の構文木を全列挙するためのアルゴリズム

自然言語データを解析する手法の一つである文脈自由文法は，文の意味解釈を数学的に定義することができ，本研究の目的にあったデータを使う体系とみなすことができる．文脈自由文法などの句構造文法を利用した機械学習は，“文法学習(Grammatical Inference)”とよばれていて，1960 年代から研究されている．与えられた語または文を文脈自由文法で解析する際には，複数個の構文解析木が得られることがあり，最適な構文解析木を選択する手法の設計は機械学習設計のポイントである．最適な解を求める方法として，深層学習などでは勾配法などを用いるが，構文解析木に対して勾配に対応する概念はまだ提案されておらず，すべての構文解析木を列挙して最適なものを選ぶことが自然な方法となる．そこで，与えられた語または文を導出するすべての構文解析木の集合に対して，研究分担者が考案した ZSDD(Zwro-supresed Sentential Decision Diagram)を用いた圧縮表現を構成するアルゴリズムを開発した．構文解析木の集合を求めてから ZSDD を用いると n の 4 乗オーダーの時間を要するが，CKY 法に沿って ZSDD を構築すれば n の 3 乗オーダーの時間で済むことを示した．

公表論文

Kei Amii, Masaaki Nishino, Akihiro Yamamoto: On Representing the Set of All Parse Trees with a Decision Diagram, Transactions of the Japanese Society for Artificial Intelligence, 34(6), pp. A-I34_1-12, 2019.

網井圭，西野正彬，山本章博：文脈自由文法による構文木の集合を表現する決定グラフの高速な構築，人工知能学会 人工知能基本問題研究会(第 105 回),および電子情報通信学会（第 166 回）アルゴリズム研究会 合同研究会，pp6-11, 2018

[2] 一階述語論理で記述された機械学習の結果を圧縮形で構成するアルゴリズム

一階述語論理は，数学で用いられる文を形式化することに端を発し，構文解析が一意に可能な文だけを対象とした上で，文の意味解釈も数学的に定義するため，本研究の目的にあったデータを使う体系とみなすことができる．そこでのデータとしては構文解析木を直接扱うと考えてよい．一階述語論理の文を対象とした機械学習は，“帰納論理プログラミング(Inductive Logic Programming, ILP)”とよばれていて，20 年以上にわたって研究されている．ILP では，学習においてデータ D だけではなく一階述語論理で表現された背景知識 H を仮定するという特徴がある．また，学習の結果を仮説ともよぶ．従来の ILP 研究では、データ D と背景知識の組 H に対して解を 1 つしか見つけられない方法を提案しているが，この解がどのような意味で最適な解なのかは手法に応じて定義されるか，不明瞭なものであった．そこでどのような最適化の基準にも使えるように，問題のすべての解を効率的に列挙するアルゴリズムを，完備束構造を利用しながら BDD(Binary Decision Diagram)を用いて再帰的アルゴリズムの形で構成した．すべての解が列挙されれば，最適化を適用するだけではなく，ユーザに好みの解を選択させる，という利用方法も可能となる．また，評価関数が与えられたときに，最適な仮説を得るための効率的な方法も提示した． ILP の研究において，BDD を利用して解を全列挙するアルゴリズムは本研究で初めて開発されたものである．

公表論文

Hikaru Shindo, Masaaki Nishino, Akihiro Yamamoto: Using Binary Decision Diagrams to Enumerate Inductive Logic Programming Solutions. ILP Up-and-Coming / Short Papers, pp.52-67, 28th International Conference on Inductive Logic Programming (ILP 2018), 2018.

新藤光，西野正彬，山本章博：二分決定グラフを用いた帰納論理プログラミングの解の列挙，人工知能基本問題研究会(第 106 回),4-19, 2018.

[3] 木構造データ間の高速な計算が可能な距離

一階述語論理においては，すべてのデータは構文解析木で表現される．そこで，木構造データ間の距離を高速な距離の計算方法として pq-gram 距離を採用し，入力として与えられる木の各部分木の重みを考慮した重み付き pq-gram 距 離を新たに定義した上で，最近傍法によって重み付き pq-gram 距離を高速に計算するアルゴリズムを開発した．木構造データ間の距離として，列構造データ間の編集距離を木構造データ間に拡張した編集距離の計算アルゴリズムを採用してしまうと，入力の木のノード数 n の 3 乗オーダーの時間を要し，しかも昨年度の本課題の成果

である列構造データ間の編集距離の計算の高速化の限界から，木構造データ間の編集距離計算の高速化には困難が伴うと予想された．そこで編集距離の代わりに，木のノード数 n に対して O(nlog n)で計算可能である pq-gram 距離を採用することとした．pq-gram 距離を機械学習に導入するにあたって，学習アルゴリズムとして最近傍法を想定し，入力として与えられる木の各部分木の重みを考慮した重み付き pq-gram 距離を新たに定義した上で，最近傍法によって重み付き pq-gram 距離を高速に計算するアルゴリズムを開発した．

公表論文

Hikaru Shindo, Masaaki Nishino, Yasuaki Kobayashi, Akihiro Yamamoto: Metric Learning for Ordered Labeled Trees with pq-grams, 24th European Conference on Artificial Intelligence (ECAI2020), Frontiers in Artificial Intelligence and Applications, 325,pp.1475-1482, 2020 (ポスター発表: 情報系 WINTERFESTA Episode 5)

新藤光，西野正彬，小林 靖明,山本章博: pq-gram を用いた木構造間の距離の学習，人工知能基本問題研究会(第 110 回),13-18, 2020.

[4] 精密化演算子を利用した論理式の探索による機械学習

一階述語論理の一部である Horn 論理を用いた機械学習に対しては，精密化を利用して仮説の候補を探索することが有用であることが，研究の初期の段階から提案されてきた．精密化とは，論理式の意味が「減少」するように論理式を変形することであり，現代的な視点で見れば論理式の「差分」を作ることに相当し，精密化を用いた探索とは一種の「降下法」と解釈することができる．そこで，精密化を用いた探索を現代的な最適化としての機械学習に適用する方法を考案した．

公表論文

Hikaru Shindo, Masaaki Nishino, Akihiro Yamamoto: Differentiable Inductive Logic Programming for Structured Examples, Proceedings of AAAI-21 Technical Tracks 6, 35, 5034-5041, 2021.

[5] 語列データの最小汎化の計算と限界

自然言語データを構文解析せずに直接扱う語列モデルについては，部分列に着目して文字列に変数を導入したパターンとよばれる一種の代数的な式を導入した上で，パターン間の半順序構造を定義した上で，最小汎化を最適な解と考えれば，最小汎化の計算アルゴリズムは最適化アルゴリズムであり，機械学習アルゴリズムとみなすことができる．このような機械学習は文法学習と帰納的論理プログラミングの中間種と考えることができる．まず 2 つの文字列の最小汎化の計算アルゴリズムを動的計画法に基づいて設計した．このアルゴリズムは，入力となる 2 つの文字列の長さの積のオーダーで動作する．また，長さが同じ n であるような k 個の文字列からなる集合の最小汎化は k×(n の k 乗)のオーダーで動作するようなアルゴリズムも設計可能である．さらに，最近提唱され注目されている強指数時間仮説 (SETH)を認めれば，これ以上の改良が困難であることも示した．

公表論文

里見 琢聞，小林 靖明，山本 章博：文字列データの線形最小汎化問題に対するアルゴリズム，人工知能基本問題研究会(第 109 回), pp.78-82, 2020.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 HikaruShindo, Masaaki Nishino, Akihiro Yamamoto	４．巻 35
２．論文標題 Differentiable Inductive Logic Programming for Structured Examples	５．発行年 2021年
３．雑誌名 Proceedings of AAAI-21 Technical Tracks 6	６．最初と最後の頁 5034-5041
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 HikaruShindo, Masaaki Nishino, Yasuaki Kobayashi, Akihiro Yamamoto	４．巻 325
２．論文標題 Metric Learning for Ordered Labeled Trees with pq-grams	５．発行年 2020年
３．雑誌名 Frontiers in Artificial Intelligence and Applications	６．最初と最後の頁 1475-1482
掲載論文のDOI（デジタルオブジェクト識別子） 10.3233/FAIA200254	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 Kei Amii, Masaaki Nishino, Akihiro Yamamoto	４．巻 34(6)
２．論文標題 On Representing the Set of All Parse Trees with a Deci-sion Diagram	５．発行年 2019年
３．雑誌名 Transactions of the Japanese Society for Artificial Intelligence	６．最初と最後の頁 A-I34_1-12
掲載論文のDOI（デジタルオブジェクト識別子） 10.1527/tjsai.A-I34	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7件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Hikaru Shindo, Masaaki Nishino, Yasuaki Kobayashi, Akihiro Yamamoto
２．発表標題 Metric Learning for Ordered Labeled Trees with pq-grams
３．学会等名 24th European Conference on Artificial Intelligence (ECAI2020)（国際学会）
４．発表年 2020年

１．発表者名
新藤光 , 西野正彬, 小林靖明, 山本章博

２．発表標題
Metric Learning for Ordered Labeled Trees with pq-grams

３．学会等名
情報系 WINTERFESTA Episode 5

４．発表年
2019年

１．発表者名
新藤光 , 西野正彬, 小林靖明, 山本章博

２．発表標題
pq-gramを用いた木構造間の距離の学習

３．学会等名
人工知能学会 人工知能基本問題研究会資料（第110回）

４．発表年
2019年

１．発表者名
里見 琢聞, 小林 靖明, 山本 章博

２．発表標題
文字列データの線形最小汎化問題に対するアルゴリズム

３．学会等名
人工知能学会人工知能基本問題研究会（第109回）

４．発表年
2019年

１．発表者名
Hikaru Shindo, Masaaki Nishino, Akihiro Yamamoto

２．発表標題
Using Binary Decision Diagrams to Enumerate Inductive Logic Programming Solutions

３．学会等名
28th International Conference on Inductive Logic Programming (ILP 2018), (国際学会)

４．発表年
2018年

１．発表者名 新藤光，西野正彬，山本章博
２．発表標題 二分決定グラフを用いた帰納論理プログラミングの解の列挙
３．学会等名 人工知能学会　人工知能基本問題研究会（第106回）
４．発表年 2018年

１．発表者名 網井圭，西野正彬，山本章博
２．発表標題 文脈自由文法による構文木の集合を表現する決定グラフの高速な構築
３．学会等名 人工知能学会　人工知能基本問題研究会（第105回）および電子情報通信学会　第166回アルゴリズム研究会　合同研究会
４．発表年 2018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
研究分担者	西野　正彬 (Nishino Masaaki) (90794529)	日本電信電話株式会社ＮＴＴコミュニケーション科学基礎研究所・協創情報研究部・特別研究員 (94305)	

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