（ 続紙 １ ）

京都大学	博士（　　情報　　学）	氏名	杉　山　麿　人
論文題目	Studies on Computational Learning via Discretization （離散化に基づく計算論的学習に関する研究）		

（論文内容の要旨）

本論文は，実数値データなどの連続的対象からの機械学習に対して，対象の離散化と機械学習の２つのプロセスを融合するという方針に基づいて，計算論的理論を構成するとともに，実問題への適用を目指したアルゴリズムを提案するものである．連続的対象からの機械学習は広く利用されているにも関わらず，そこで対象の離散化が陽に意識されることはほとんどない．本論文は，このような背景を念頭に置きつつ，離散化を取り込んだ機械学習について，理論と具体的アルゴリズムを提案している．

第１章は序章であり，実数値データの離散化と計算理論および学習理論について述べ，続いて本論文の概要を示している．

第２章では，極限同定モデルに基づき，ユークリッド空間上のコンパクト集合の学習を解析することで，実数値データに対する２値分類問題の計算論的基礎を与えている．従来の極限同定学習の解析方法に倣い，様々な学習基準における学習可能性の階層を明らかにし，学習の複雑さをHausdorff次元とVC次元により定量化することで，機械学習とフラクタル幾何との関係を示している．さらに，計算可能性解析学における実数計算のタイプ２理論に基づいて，計算可能性の側面から学習を解析している．

第３章では，２つの実数値データ集合間の異なり具合を測るために，符号化ダイバージェンスという新規の計算論的な尺度を提案している．実数の符号化を利用して，２つのデータ集合を分離する際の困難さを表現しており，この尺度を用いた怠惰学習器が，従来のクラス分類手法に遜色ない性能を持つことを示している．

第４章では，符号化ダイバージェンスを応用することにより，クラスタリングの結果を評価するための新規の基準として最小符号長(MCL)を提案している．さらに，MCLを最小化するクラスタリング・アルゴリズムCOOLとG-COOLを提案している．COOLでは２進符号化を用いることにより，その実行時間を入力データ数と次元数の積のオーダにしている．一方，G-COOLでは，Gray符号化を用いることにより，任意形状のクラスタを抽出することを可能としている．

第５章では，前章の結果を発展させ，任意形状のクラスタを抽出し，かつ高速であるという２つの要求を満足するクラスタリング・アルゴリズムであるBOOLを提案している．そこでは，ソーティング・アルゴリズムを応用することで２つの要求を満たすことに成功している．計算機実験によって，BOOLは，著名な従来アルゴリズムであるK-meansアルゴリズムよりも高速に動作することを確認し，さらに，任意形状のクラスタを抽出可能なアルゴリズムとして他研究が提案したものよりも100～1000倍高速に動作することを示している．

第６章では，半教師あり学習と順序学習というさらに複雑な機械学習のタスクに取り組んでいる．形式概念解析を利用することにより，離散値と連続値が混在するデータから，クラスの分類とランキングを半教師あり学習によって達成するアルゴリズムSELFを提案し，その有効性を実験的に示している．

第７章では，生物学データへの応用に取り組んでいる．生物学データベースからの

リガンド候補の発見を半教師ありのマルチ・ラベルクラス分類問題として定式化した上で，その問題を解くアルゴリズムLIFTを構築している．さらに，このアルゴリズムの有効性を実験的に示している．

第８章では，本論文で与えた結果を概観し，結論としている．

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は，実数値データなど連続的な対象からの機械学習に，離散化のプロセスを組み込むことにより，計算論的機械学習理論と新たな機械学習アルゴリズムを構成することを目的としている．現在，機械学習は様々な分野で利用されているが，理論上は連続的データを仮定しているにも関わらず，データの蓄積や学習アルゴリズムの実装は離散化されたデータを用いて行われるという隔たりがある．本論文は，この問題点を解決するための理論を与え，その考察に基づいて健全かつ実効的で効率的な機械学習アルゴリズムを与えている．主要な結果は以下の３つである．

１．実数値データに対する２値分類問題の計算論的基礎を与えることを目的とし，計算論的学習における基本的なモデルの１つである極限同定学習に基づいて，離散化を組み込んだ新たな学習モデルを構築した．さらに，このモデル上で学習に様々な条件を課した場合の学習可能性を明らかにし，それを階層の形で提示した．また，学習の複雑さをHausdorff次元とVC次元を用いて定量化することに成功し，機械学習とフラクタル幾何との関係を明らかにした．

２．連続的な対象の離散化が，自然にクラスタリングのプロセスとみなすことができるという着想に基づき，符号化ダイバージェンスという実数値データ集合間の異なりを測るための新規の尺度を提案した．これは，機械学習において現在主流となっている統計的アプローチとは一線を画すものである．この尺度に基づく怠惰学習によるクラス分類の能力を検証している．さらに，前述の着想を発展させることにより，新たなクラスタリング評価基準を提案した上で，高速かつ柔軟なクラスタリング・アルゴリズムを構築した．その結果として，Gray符号化が任意形状のクラスタの抽出に寄与することを示した．さらに，この結果を改良することにより，任意形状のクラスタを抽出可能なアルゴリズムとしては世界最速であるものを構築することに成功した．

３．形式概念解析と呼ばれる代数的データ解析手法と離散化を組み合わせることにより，複雑な機械学習のタスクとして現在注目されている半教師あり学習と順序学習を達成するアルゴリズムを構築した．このアルゴリズムは，離散値と連続値が混在しているデータに対して直接適用可能である．さらに，このアルゴリズムを生物学的データからの知識発見に適用するために改良し，その結果得られたアルゴリズムが従来アルゴリズムと同等もしくはそれ以上の性能を持つことを実験的に示した．

これらの研究成果は，どれも全く新規な着想に基づいて行われている．今日，統計的アプローチによる機械学習アルゴリズムが普及しているが，本論文では，離散化を組込むことによって，計算論的アプローチでも様々な学習が達成できることを理論と実践の両面から示しており，その独創性は特筆すべきである．また，既存のものと同等以上の性能を持つアルゴリズムを構築しており，機械学習分野に対する貢献度は高い．よって，本論文は博士（情報学）の学位論文として価値のあるものと認める．

また，平成２４年２月２０日に実施した論文とそれに関連する内容についての口頭試問の結果，合格と認めた．