学位論文題名

Learning and Detecting Concept Drift

(Concept Drift の学習と検出に関する研究)

学位論文内容の要旨

情報科学の発展によって人類が創り出すデータ量は爆発的に増大している。このため、大量のデータから入出力関係や知識表現を抽出する機械学習の需要が高まり、現在では高度情報化社会を支える基盤技術の一つにまで成長した。音声・文字認識、検索エンジン、医療支援、ロボット工学、画像処理、データマイニングなど、これまでに機械学習の技術が応用され大きな成果を挙げた分野は多岐に渡る。近年では、計算機の記憶容量や演算処理能力の飛躍的な向上を背景に、過去の計算機では実現不可能だった手法が次々と開発されており、今後も機械学習のさらなる発展が予想される。

機械学習を訓練サンプルの与え方によって大別すると、まとめて与えられた大量のサンプルを一度に学習するバッチ学習と、順次与えられるサンプルを追加的に学習するオンライン学習の二種類に分けられる。このうち、バッチ学習については非常に優れた手法が多数提案されているが、オンライン学習、特に、学習対象の基となる統計的な性質が時間と共に変化する環境下での学習については多くの課題が残されている。この変化は *concept drift* と呼ばれ、緩やかな変化のみならず、突然かつ大きな変化をも含む。変化の性質によって学習システムが採るべき戦略は大きく異なるため、様々な性質の変化を含む問題の解決は非常に難しい。その実例の一つとしてはスパムフィルタリング問題が挙げられる。この問題においてスパムフィルタは、ユーザが持つスパム判定基準の僅かな揺らぎやメールで扱う話題の変化、さらには狡猾なスパム送信者が起こす変化など、日々発生する様々な性質の変化に対応しながら学習を進めなければならない。

今後は様々な分野において、幾多の変遷を経ながら大量に創出され続けるデータを処理するために、逐次的・追加的に学習可能な、そして高い順応性を持ったシステムがより一層必要となるであろう。そこで本論文では、与えられた入力サンプルに対応するクラスを予測するクラス分類問題全般を対象に、*concept drift* に対応可能なオンライン学習システムを様々なアプローチから複数提案する。全ての提案システムは *concept drift* の発生を検出して学習に利用するという特徴を持つ。

以下に、全6章から成る本論文の構成を示す。

第1章では、導入として研究背景、目的、本論文の構成について述べる。

第2章では、基本原理として既存のバッチ学習システムとオンライン学習システムを示す。

第3章では、複数のクラス分類器を使用する提案システム ACE (*adaptive classifiers-ensemble*) について示す。複数のクラス分類器を用いるシステムは、単一のクラス分類器を修正しながら使用するシステムよりも *concept drift* に対応し易い場合が多く、近年多くの研究が行われている。ACE は

システム全体の出力を、常に更新を行う一つのオンラインクラス分類器と、作成後は更新を行わない多数のバッチクラス分類器の出力の重み付き多数決によって決定する。そして、各クラス分類器の最近のサンプルに対する予測精度とその信頼区間を利用して concept drift の検出を行い、新たなバッチクラス分類器を順次追加していく。ACE は再発する変化に対して従来手法よりも高い性能を実現するが、変化検出の精度とシステム構成の複雑さに問題があった。そこで、以降の研究においては、concept drift の正確な検出を単純なシステムによって実現することを目指した。

第4章では、二つのオンラインクラス分類器を利用して concept drift の学習と検出を行う提案システム Todi (*two online classifiers system for learning and detecting concept drift*) について説明する。まず、我々は新たな変化検出法として STEPDP 手法を提案した。STEPDP は単一のオンラインクラス分類器についてその予測精度を監視し、concept drift の発生が引き起こす最近の予測精度の急激な悪化を統計的検定 (*a statistical test of equal proportions to detect concept drift*) によって検出する。そして、検出後はクラス分類器を初期化して新たな学習に備える。STEPDP は ACE で使用する手法や他の従来手法に比べ高い検出性能を実現したが、誤検出が発生した場合にシステムの精度が大きく悪化する問題があった。そこで、誤検出が発生しても予測精度が悪化しない学習システムの実現を目指して我々は Todi を提案した。Todi は、変化検出後に初期化されるクラス分類器に加えて、検出後も継続して学習を行うクラス分類器を使用する。この工夫によって、Todi は高い予測精度を常に維持し続けると共に、過去の検出の正誤を自ら判断して concept drift の発生をユーザに正しく通知できる様になった。Todi の有効性を検証するためスパムフィルタリング問題に取り組んだところ、高性能なスパムフィルタとして知られる Bogofilter をオンラインクラス分類器として二つ使用した Todi は、単体の Bogofilter よりも高い予測精度を実現した。今後は ACE と Todi の技術を結合して、多くの実世界問題を解決して行きたい。

第5章では、人間が行う変化検出に関する知見を基にした提案手法 LID (*leaky integrate-and-detect*) について示す。我々は、人間が行う変化検出と、どのような誤分類も全て同様に扱って変化を検出する機械学習手法との間には大きな隔たりがあると考えた。そこで人間の変化検出について調査し、認知科学的知見を機械学習に取り入れることを目指して研究を行った。まず「最近の予測精度が高い状態で確信度の高い回答が否定されるほど、人間は変化を高速に検出できる」という作業仮説を立てて行動実験を行い、仮説の正しさを示唆する結果を得た。そして、自信のある回答が連続して否定されることの重要性を発見した。これらの知見から、スパイク発生機構の単純なモデルである leaky integrate-and-fire モデルを変化検出に応用できると考え、これを基にした手法 LID を提案した。LID は単一のオンラインクラス分類器が持つ確信度と最近の予測精度を利用して、突然かつ重大な変化を検出する。その有効性を検証するため計算機実験を行ったところ、ノイズや緩やかな変化が存在する環境下でも、LID は高速かつ正確に変化を検出できた。今後は本研究で得た確信度に関する知見を Todi や ACE に導入して行きたい。

第6章では、結論として本論文の成果と今後の課題を示す。

以上をまとめると、我々は学際的な研究を行って、concept drift の学習と検出が可能なオンライン学習システムを複数提案した。本論文は concept drift を扱うオンライン学習システムが持つべき能力の実現に貢献を果たした。

学位論文審査の要旨

主査 准教授 山内 康一郎
副査 教授 栗原 正仁
副査 教授 大内 東
副査 教授 和田 充雄
副査 教授 大森 隆司 (玉川大学脳科学研究所)

学位論文題名

Learning and Detecting Concept Drift

(Concept Drift の学習と検出に関する研究)

近年のインターネットの発達によって、コンピュータは大量のデータを扱えるようようになった。しかし、その一方で、洪水のように押し寄せる大量の情報をそのまま扱うのは現実的ではなく、大量のデータから規則性を抽出して利用する技術が求められている。その一つのアプローチとして入力関係や知識表現を抽出する機械学習を利用する方法があり、近年その重要性が高まっている。

学習機械の学習方法には、大きく分けて二通り存在する。一つは、バッチ学習法であり、大量のサンプルをあらかじめ貯めておき、一度に学習する方法である。もう一つはサンプルを逐次的に学習するオンライン学習法である。実用上は、全てのデータをあらかじめため込む必要のないオンライン学習法が有効である。

しかし、このオンライン学習法には、様々な課題が残されている。特に学習対象の統計的な性質が時間と共に変化する状況での学習は困難を極める。一般にこのような変化は **concept drift** と呼ばれており、従来型の学習理論の多くがその礎とする、静的環境を前提とする統計理論が成立しない。これに対処するためには、**concept drift** の発生を検知し、その時点で学習機械の内部パラメータをリセットするか、もしくは再発する **concept** の場合には、過去に獲た学習結果を再利用する等の対策が必要となる。つまり、当該領域ではできるだけ正確な **concept-drift** 検出手法を構築することが極めて重要である。

ところで、従来にも **concept-drift** を扱う学習理論はいくつか提案されてきた。それらの多くは、緩やかな変化のみへの対処方法であったり、もしくは突然かつ大きな変化のみへの対処方法であった。これは、それぞれの環境下に即した学習戦略が、お互いに異なるためである。しかし実環境では、緩やかな変化と急激な変化の両方が生ずるものである。したがって、この両方の **concept drift** に対処できる学習機械が必要である。

このような状況をふまえ、本論文は、緩やかな **concept drift** と急激な **concept drift** との両方に対処するオンライン学習システムを実現することを目的として、そのための学習法と変化検出法とをいくつか提案している。特に変化検出法については重要視しており、大きく分けて三つの手法を提案し

ている。ただしここでは、与えられた入力サンプルに対応するクラスを予測するクラス分類問題全般を対象としている。

本論文の第1章では、研究背景、目的、本論文の構成について述べている。

第2章では、既存のバッチ学習システムとオンライン学習システムを示している。

第3章では、複数のクラス分類器を使用するシステム ACE について提案している。ACE はシステム全体の出力を、常に更新を行う一つのオンラインクラス分類器と、作成後は更新を行わない多数のバッチクラス分類器の出力の重み付き多数決によって決定する。そして、各クラス分類器の最近のサンプルに対する予測精度とその信頼区間を利用して **concept drift** の検出を行い、新たなバッチクラス分類器を順次追加していく。ACE は再発する変化に対して従来手法よりも高い性能を実現している。

第4章では ACE を単純化して、**concept drift** 検出精度向上に焦点を絞った手法を提案している。まず、新たな変化検出法として STEPDP 手法を提案している。これは単一のオンラインクラス分類器についてその予測精度を監視し、**concept drift** の発生が引き起こす最近の予測精度の急激な悪化を統計的検定によって検出する。**concept drift** が検出されると、そのオンラインクラス分類器は初期化される。STEPDP は、ACE で使用する手法や他の従来手法に比べ高い検出性能を実現するが、誤検出が発生した場合には学習器が誤ってリセットされるため、システムの精度を大きく落としてしまう。この欠点を克服するため、二つのオンライン分類器を使用して、誤検出が起きても学習精度の劣化しない Todi システムへと発展させている。この手法は、変化検出後に初期化されるクラス分類器に加えて、検出後も継続して学習を行うクラス分類器とを使用する。この工夫によって、Todi は高い予測精度を常に維持し続ける。計算機実験では、bogofilter と呼ばれる分類器を二つ利用した Todi をスパムフィルタに適用している。その結果、現在多用されている単体の bogofilter によるスパムフィルタよりも良い結果が得られることを示している。

第5章では、人間が行う変化検出に関する知見を基にした提案手法 LID について示している。人は時として学習機械よりも驚くほど素早く状況の変化を検出できることを指摘し、人間の変化検出法を **concept drift** 検出に応用することを目指している。このために著者は「最近の予測精度が高い状況で確信度の高い回答が否定されるほど、人間は変化を高速に検出できる」という作業仮説を立てて行動実験を行っている。そしてその結果を基に、スパイク発生機構のモデルである leaky integrate-and-fire モデルを変化検出に応用した手法 LID を提案している。LID は、単一のオンラインクラス分類器が持つ確信度と最近の予測精度とを利用して、突然かつ重大な変化を検出する。計算機実験によって、ノイズや緩やかな変化が存在する環境下でも、LID は高速かつ正確に変化を検出できることを示している。

第6章では、結論として本論文の成果と今後の課題を示している。

以上のようにこの研究では **concept drift** を含む環境での学習に必要な手法をいくつか提案したものとなっている。特に変化検出手法に関する手法に関しては重きを置いて議論されている。今後ここで提案された手法が組み合わせられ、様々な **concept drift** 環境に対応する学習システムの構築がなされることが期待される。

これを要するに、著者は、オンライン学習、特に、学習対象の基となる統計的な性質が時間と共に変化するという **concept drift** が引き起こす問題について、**concept drift** の学習と検出の方法の新知見を得たものであり、情報科学、特に機械学習分野の発展に貢献するところ大なるものがある。よって著者は、北海道大学博士(情報科学)の学位を授与される資格あるものと認める。