

< 修 士 論 文 >

観測にバイアスを持つ状況下
における二値分類問題
(要 旨)

滋 賀 大 学 大 学 院
デ ー タ サ イ エ ン ス 研 究 科
デ ー タ サ イ エ ン ス 専 攻

修了年度：2022年度

学籍番号：6021142

氏 名：渡邊翔太郎

指導教員：松井秀俊

提出年月日：2023年1月10日

要旨

昨今の計測・測定技術の向上により、様々な分野で多種多様なデータを得ることが可能になり、様々なデータの背景に対して適した手法を当てはめることが必要とされている。二値分類の目的は、サンプルが正のクラスか負のクラスのどちらに属するか識別することである。二値分類問題において、一般的な教師あり分類では、正のデータと負のデータを用いて分類器を学習していた。ここで、正のデータと負のデータは、それぞれ手元にあるデータから、それらを生成している分布の全体を表現できるものであり、このことから、一般的な教師あり分類は厳しい仮定を必要としているといえる。これに対して、本論文では、負のデータが多様であり包括的に特徴つけることができない状況や、それ以前に、負のデータの大多数が得られない状況を考える。この状況において、一般的な教師あり分類で必要としているような負のデータを得ることは困難だが、ある特徴に偏った一部の負のデータを得るだけなら比較的容易な場合がある。本論文では、正のデータと、そのような観測に偏りのある負のデータから、二値分類器を学習する弱教師あり学習法である、PbN (Positive and biased Negative) 分類問題にアプローチする新しい方法を提案する。

しかし、PbN 分類では、観測されているデータである事後確率を正確に求められない問題がある。そこで、この問題に対するアプローチとして、本研究では観測されているデータが正である事後確率を表す信頼度の歪みによる悪影響を修正する方法を組み込む。ここで、本論文では信頼度などがバイアスによって変形することを「歪み」と称する。これにより、PbN 分類問題の経験リスク最小化に必要な、観測されているデータである事後確率が歪んでしまう影響を軽減することができる。本論文では、数値実験と実データ解析により、提案手法と観測されているデータである事後確率の歪みを、修正する前の PbN 分類、一般的な教師あり分類を比較することで提案手法の有効性を検証する。

本論文は5節と1つの付録から構成される。全体の流れは以下になる。まず、本研究の背景と目的を説明し、課題とそれに対する解決策を述べる。二値分類問題において、一般的な教師あり分類で想定している背景は非常に限定的であることを述べ、その背景を持たない様々な背景のデータに対するアプローチについて概観する。そして、それらの背景の一つとして、観測に偏りのある一部の負のデータが得られる場合に着目し、分類器を構築する上での課題とそれに対する解決策を述べる。次に、本研究に関連する一般的な教師あり分類と様々な背景設定に対する分類手法の問題設定について説明する。具体的には、正のデータとラベルのないデータから二値分類を行う PU 分類、正のデータとラベルのないデータと偏りのある一部の負のデータから二値分類を行う PUbN 分類、信頼度を与えられた正のデータのみから二値分類を行う Pconf 分類について説明する。

続いて、PbN 分類のための分類リスクを提案する。ここでは、観測されているデータであれば+1、未観測の負のデータであれば-1 を返す潜在変数を導入し、PbN 分類リスクを導出した。しかし、観測されているデータである確率は負のデータを生成している分布の全体

に依存しているため陽に求められず、疑似的な分布を利用することで、歪んだ観測されているデータである確率をかろうじて求めるに留まっており、分類精度に悪影響を及ぼす可能性がある。そこで、その歪みを軽減するために、観測されているデータが正である事後確率を表す信頼度の歪みを軽減する方策を組み込んだ。ここでは、偽陰性率が既知であるという仮定を置き、既知である偽陰性率と経験的偽陰性率の差が最も小さくなるように観測されているデータである確率をハイパーパラメータで調整することで、分類精度の改善を図った。また、あくまでも観測されているデータである確率に歪みが生じない場合に限定されるが、PbN 分類に対する理論解析を行った。

以上を踏まえて、提案手法の性能を人工データとベンチマークデータを用いての実験で検証した。人工データを用いての実験では、負のデータの散らばり具合と、得られたデータの偏り方に着目し、4 通りの状況下で実験を行った。また、既知と仮定した偽陰性率にずれがあった場合や高次元に拡張したときの提案手法の分類精度に与える影響を検証した。ベンチマークデータを用いての実験では、無線による屋内測定データセットと MNIST データセットを使用した。付録では、様々な分類手法の分類リスクと、提案手法の理論解析に必要な補題や定理の導出を確認する。