審査の結果の要旨

氏　名　　宮口　航平

大量のデータからの知識を獲得するための機械学習において、近年では高次元モデルを扱うことが多くなってきている。高次元モデルとは、ニューラルネットやグラフィカルモデルのようなパラメータの数が非常に多いモデルのことである。このようなモデルを学習する場合、過学習を起こしやすくなったり、計算量が膨大になったりするところが問題であった。本論文では、上記の問題を克服するために、記述長最小化原理(Minimum Description Length (MDL) principle) の意味で最適な高次元モデルを効率的にするための方法論を開発した。MDL原理とは、与えられたデータに対して総記述長を最小にするモデルを最適なモデルと見なすモデル選択原理である。提案する方法論の核となるのは、MDL原理の意味で最適な高次元モデルを（1）緩和的かつ確率的に解く方法、（2）解析的かつ近似的に解く方法、（3）ベイズの周辺確率モデルとして近似的に解く方法である。これらの方法を実現する具体的アルゴリズムを構築するとともに、理論的な性能保証と実験的検証を行うことにより、高次元モデルの学習の体系を築いた。

本論文は「Learning High-dimensional Models with the Minimum Description Length Principle (記述長最小化原理による高次元モデル学習)」と題し、7章からなる。

第1章「Introduction」では、機械学習の一原理としてのMDL原理を説明し、高次元モデルを学習することについての問題を提起している。

第2章「Preliminary」では、MDL原理に基づく機械学習の目的関数として、基本的な概念である正規化最尤符号長 (Normalized Maximum Likelihood Codelength : NML) についての概要を示している。

第3章「Graphical Model Selection via Relaxed Stochastic Complexity」では、グラフィカルモデルのような高次元スパースなモデルを対象とし、これをMDL原理の意味で最適なモデルを緩和的かつ確率的に計算するための方法を提案している。MDL原理は確率的コンプレキシティという量（正規化最尤符号長に等価）を最小化するものであるが、そこに連続的なパラメータを導入して双対問題を考え、連続最適化の問題としてこれを解く方法を与えた。その際の目的関数となる量としてRelaxed Stochastic Complexity (RSC) という量を提案した。RSCを最小化するような学習を行うための確率勾配法に基づくアルゴリズムを与え、これを高次元スパースなグラフィカルモデルの学

習に適用した。その結果、LASSOと呼ばれる手法に交差検証法を組み合わせて得られる従来手法に比べ、真のモデルの推定精度及び記述長の観点から提案手法が優れていることを経験的に示した。

第４章「High-dimensional Penalty Selection via Analytic Approximation of Minimax Regret」では、ある制約された状況下で、MDL原理の意味で最適な高次元モデルを解析的かつ近似的に計算するための方法を提案している。MDL原理の目的関数となる正規化最尤符号長において、パラメータの事前分布をおいて修正したものは、Luckiness Normalized Maximum Likelihood Codelength (LNML符号長) と呼ばれている。この量は解析的に計算することは従来困難であったが、対象とするモデルにある滑らかさの制限を置くことにより、LNML符号長の上界値を解析に計算する手法を与えた。この上界値がタイトなものであること、及びこの上界値を最小化するモデルの収束性を理論的に示した。また、線形回帰モデルなどを用いて実験的にもその性質を確認した。

第５章「Minimax Regret for Smooth Logarithmic Losses over High-Dimensional $\ell 1$-Balls」では、高次元モデルを用いたオンライン予測問題を考え、MDL原理の意味での最適戦略をベイズの周辺確率モデルとして近似的に計算する方法を提案している。MDL原理の意味での最適戦略をベイズの周辺確率モデルで近似する方法は、高次元ではないモデルについては既に知られており、Jeffreysの事前分布を使えばよいというものであった。しかしながら、データの数よりも次元の方が十分に大きいような高次元モデルでは、その方法では最適性が成り立たない。そこで、新たにSpike and Tail事前分布に基づくベイズの周辺確率戦略を提案した。また、この戦略の予測性能を評価するためのツールとしてEnvelop Complexityという概念を提案した。これを用いて、上記戦略に対する累積予測損失指標であるリグレットという量の上界を与え、これが下界と漸近的に定数倍の差に収まることを理論的に示した。

第６章「Excess Risk Bounds with Envelope Complexity」では、第５章と同じく、オンライン予測問題を考え、あるパラメータ値を代入して実現する戦略を提案している。そのパラメータ値はギブスの事後分布に従って得られるもので、従来よりPAC-Bayesの方法として知られてきた。その戦略のリグレットについて、第5章で提案したEnvelop Complexityに焼きなまし法を用いることにより、その上界を与えた。

第7章「Conclusion」では全体を総括し、将来の展望を与えている。

以上を要するに、MDL原理に基づく学習を高次元モデルに適用する際の計算上の問題を克服し、理論的な推定・予測性能の限界を評価する方法論を構築している。高次元モデルの学習という問題設定においては、MDL原理に基づく学習の適用方法や性能解析はこれまで研究されていなかった。本論文は、こうした新しい問題に対して、有効な方法論を体系的に提示しており、数理情報学の発展に大きく寄与している。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。