「過去の長期気候変化シグナルの検出とその要因推定」野沢 恭(国立環境研究所)
「成層圏突然昇温現象の予測可能性」向川 均(京都大学防災研究所)
「地球温暖化予測の不確実性と確率論的アプローチ」江守正多(国立環境研究所)
「地球温暖化による健康影響」小野雅司(国立環境研究所)

【オーガナイザー】
江守正多、野沢 恭(国立環境研究所)
椿 広計(筑波大学大学院、統計数理研究所)
松本幸雄(社)国際環境研究協会
金藤浩司、河村敏彦、友定充洋(統計数理研究所)

研究紹介
統計的予測問題におけるアンサンブル学習の有効性
数理・推論研究系 伏木 忠義

既存のデータをもとに将来のデータの値を予測する方法は、日常で広く必要とされる技術である。このような予測問題に対して、統計学や機械学習など、さまざまな分野で研究が行われてきた。予測問題は基本的な問題であり、長年の研究があるが、使える計算リソースや数値問題の大きさ、データ量といった要因は、時代と共に変化し、それに対応した手法が提案、研究されている。

機械学習の分野では、アンサンブル学習とよばれる手法が研究されている。アンサンブル学習は、何個も学習を行い、その結果得られた多数の学習機械をうまく組み合わせて予測を行うものである。アンサンブルをとることで1つの学習機械のみで予測するよりも良い予測が実現される場合がある。特に、ブートストラップデータを用いて何個も学習を行い、その結果の単純平均をバッシングや学習がうまくいかなかった例題について重みをつけて次回の学習を行うブースティングという手法が有名である。一方、統計学においても、Kullback-Leibler ダイバージェンスを損失関数とした統計的予測問題の文脈で、推定量をパラメータのところに代入したプラグイン予測、ブートストラップデータを用いた予測、Bayes 予測など予測方法についてこれまで研究されている。

統計学で提案されたブートストラップ予測は、学習理論のアンサンブル学習の立場からは、最尤推定量のプラグイン分布にバッシングを適用したものと考えることができる。Bayes 予測は、事後分布からパラメータをサンプリングして平均をとったものと考えることができる。特にアンサンブル学習の1つの目安である。

私のこれまでの研究では、このように2つの異なる分野で研究されていた手法を統一的な観点でとらえ、統計的予測問題におけるアンサンブル学習の効果について調べてきた。

まず、真の分布が仮定した統計モデルに含まれている場合を考えよう。このときは、Bayes 予測が許容的となるため、Bayes 予測がある意味で最良の予測を与えるといえる。私の研究では、最近近似を用いて、ブートストラップ予測が Bayes 予測の近似になっていることを示し、さらに予測性能に関しては、ブートストラップ予測が最尤推定量のプラグイン分布よりも最近近似に良い予測を与えることを示した。

一方、統計学においては、統計モデルに真の分布があると仮定して議論する場合が多いが、現実的には、複雑な現象を扱う場合など、そのような仮定が成立立つことは稀であると考えられる。仮定した統計モデルの中に真の分布に十分近い分布が存在すれば問題ないが、そのような仮定が成立立つことは一般には保障できないだろう。

真の分布がモデルに含まれる場合には、ブートストラップ予測は Bayes 予測の近似と考えられがた、私の研究では、真の分布がモデルに含まれない場合は、ブートストラップ予測と Bayes 予測には違いがあるということがわかった。特に、予測性能については、ブートストラップ予測が無情報の Bayes 予測よりも最近近似には良い予測を与えることを示した。

アンサンブル学習は、繰り返しの学習を必要とするため、大きな計算リソースを要求する。計算機性能の向上にとどまない、ある程度の計算コストがかかったとしても予測が欲しいという状況はあろう。アンサンブル学習は、このような状況で
有効な手法であり、私の研究もその意味で、現代的な観点で意味がある研究といえる。

参考文献