巻頭言

AIとビッグデータ

京都工芸繊維大学 情報工学･人間科学系 教授　寶　珍　輝　尚

近年、AI（人工知能）が脚光を浴びている。2011年にIBMのワトソンが人間のクイズ王に勝利し、2015年にはGoogle DeepMindが開発したAlphaGoが人間のプロ囲碁棋士を破り、2016年にはIBMのワトソンが診断の困難な白血病の診断を行い適切な診断を行ったことが報じられている。この背景には、計算機の性能が劇的に向上し、計算機で大量のデータを利用し、処理することが可能になってきたことがある。いわゆるビッグデータと呼ばれる非常に大量の時変データの活用が注目を集めてきたのもこの頃である。

AIと呼ばれているのは、人工知能分野と計算知能分野の機械学習や伝統的な統計解析手法、または、機械学習後のモデルを組み込んだプログラムを指すことが多い。機械学習も学習であり、学習次第でどうにでもなる。最初は同じでも、男女を見分けられるように学習したAIと老若を見分けられるように学習したAIは、当然ではあるが、見分けられるものが異なる。また、学習の仕方によって成功することもあれば失敗することもある。どのような学習をしたのかが重要であり、AIを使えば何でもうまくゆくということはないことが分かる。

では、どうするとAIを成功に導けるのであろうか。まずは、データの数である。1入力からの1出力の推定を考えよう。入力と出力の関係が直線で表されるとすると、直線の傾きと切片を求めれば良く、相異なる2点があれば良い。しかし、この2点に誤差が含まれると、2点のみから求まった傾向直線が正しいとは限らない。ここで良く利用される方法が、多数の点を使用し、これらの点からのずれを最小にする直線を求めるという方法である。この方法では平均や分散を使用することになるので、意味のある平均や分散が求められる数の点を使用する必要がある。一般に、未知数の10倍の点（データ）を使用すると良いと言われており、今の場合、未知数が2なので20個の点を使用すると良いことになる。機械学習では、内部で未知の変数を学習により決定している。これまでと同じ考え方が適用できるとすると、機械学習の内部で使用している変数が20の場合は200個のデータが必要になる。今流行りのディープラーニング（深層学習）は、基本的にはニューラルネットワークである。8入力、1出力で100ノードの隠れ層を持つ単純なニューラルネットワークでさえ1,000以上の変数を決定しなければならない。これまでの考え方に従うと10,000個以上のデータを使用しなければならないことになる。どのような方法を使用するかによって準備するデータ数が異なるので注意が必要である。

次に、データの質である。業務等で得られるデータはきれいなものとは限らず、欠損値や外れ値がある。これらの処理もないがしろにはできない。これらの処理如何で機械学習の結果が変わってくるからである。しかし、これらの処理には困難を極めることが多い。また、データの偏りも無視できないことが多い。これらの前処理を適切に行うこともAIで成功するには避けて通れないものである。

最後に、AIの使い方である。最近はAIの様々なツールや環境が整ってきており、どれを使用して良いか戸惑うことも多い。また、何か入力すれば何か値が出る。正しい使い方をしているか、入力データは適正かを確認して利用しないととんでもないことになりかねない。注意が必要である。

新型コロナウイルス感染症のニュースの陰に隠れてAIは話題に上らなくなっているが、AIが感染症拡大防止に役立つことを願うとともに、今後の企業活動の大きな力となることを願うものである。