研究員の眼

「空振り」と「見逃し」どちらが問題?

機械学習の評価尺度には人間の感覚が生かされている

保険研究部 主席研究員　篠原　拓也
(03)3512-1823　tshino@nli-research.co.jp

現在の人間社会では、情報をもとに、さまざまな判断がなされている。

例えば、誰しもが日々行っているであろうこととして、パソコンやスマートフォンに送られてくるメールのうち迷惑メールを仕分けて削除等の処理を判断する。気象情報では、気象庁の予報官が、台風などの低気圧の接近に伴って、線状降水帯による大雨の呼びかけを行うか判断する。医療の現場では、医師が、コロナウイルスやHIVのような感染症について、検査にもとづく診断を行う、といったことが挙げられる。

これらの判断は、基準を明確化すれば機械的に行うことが可能だ。つまり、AIによる判断が可能となる。ただし、人間が行うにせよ、AIが行うにせよ、判断が100%正しいとは限らない。

線状降水帯の発生予報で言えば、発生する予報を出したのに実際は発生しなかった「空振り」と、発生する予報を出さなかったのに実際は発生した「見逃し」の、2種類の誤りが起こり得る。

今回は、こうした判断の誤りについて、AIの機械学習を含めて考えてみたい。

◇ 迷惑メールの判断は慎重に行うべき

何か将来のことを予測したり、未知のことを判断したりすることと、それが実際はどうだったかということの間には、4つの場合が考えられる。

例として、迷惑メールの判断と処理について考えてみる。「迷惑メールだと判断して削除し、実際に迷惑メールだった」。「迷惑メールだと判断して削除したが、実際は迷惑メールではなかった」。「迷惑

メールではないと判断して削除しなかったが、実際は迷惑メールだった」。「迷惑メールではないと判断して削除せず、実際に迷惑メールではなかった」の 4 つの場合があり得る。

これらのそれぞれの場合の発生数をまとめたものは、混同行列と呼ばれる。100 個のメールについて、混同行列を作ると、次表のようなものとなる。

迷惑メールの判断の混同行列

	迷惑メールと判断	迷惑メールではないと判断	計
実際は迷惑メールだった	40	8	48
実際は迷惑メールではなかった	7	45	52
計	47	53	100

この表の場合は、判断内容と実際の状況が一致していたメールが 85 個（＝40 個＋45 個）となっており、まずまずの判断結果だったといえるだろう。

残りの 15 個のメールについては、「迷惑メールだと判断して削除したが、実際は迷惑メールではなかった」誤りが 7 個。「迷惑メールではないと判断して削除しなかったが、実際は迷惑メールだった」誤りが 8 個あった。

どちらの誤りが問題だろうか。迷惑メールは確かに“迷惑”だが、それを放置したからといって直ちにパソコンやスマートフォンの使用に支障が出る訳ではない。つまり、後者のケースはそれほど深刻な問題ではないと言える。

一方、前者のケースでは、迷惑メールではないのに誤って削除してしまったことになる。もし、そのように削除したメールのなかに重要なものが含まれていたら、もしかすると取り返しのつかない事態に陥ってしまうかもしれない。つまり、前者のケースは重大な問題と考えられる。

上表の迷惑メールの判断の混同行列では、前者の 7 個の誤りを減らすために、迷惑メールの判断は慎重に行うべき、ということになるだろう。

◇ 感染症の診断は積極的に行うべき

次に、感染症の診断について見てみよう。先ほどと同様に 100 人の受検者について混同行列を作ってみたところ、次表のようになったとする。

感染症の診断の混同行列

	感染症と診断	感染症ではないと診断	計
実際は感染症に罹患していた	20	10	30
実際は感染症に罹患していなかった	5	65	70
計	25	75	100

この表の場合は、診断内容と実際の罹患の有無が一致していた受検者が 85 人(＝20 人＋65 人)だった。

残りの 15 人の受検者については、「感染症と診断されたが、実際は感染症に罹患していなかった」誤りが 5 人。「感染症ではないと診断したが、実際は感染症に罹患していた」誤りが 10 人だった。

統計学や疫学では、前者の 5 人は「偽陽性」、後者の 10 人は「偽陰性」と呼ばれる。ちなみに、「感染症と診断され、実際に感染症に罹患していた」20 人は「真陽性」。「感染症ではないと診断され、実際に感染症に罹患していなかった」65 人は「真陰性」と呼ばれる。

それでは、感染症の診断の場合は、「偽陽性」と「偽陰性」のどちらの誤りが問題だろうか。偽陽性の受検者は、実際は感染症に罹患していないにもかかわらず、自宅や医療施設などで隔離期間を過ごすことになる。当人にとっては、いい迷惑だろう。また、偽陽性の人が多数入院すれば、医療資源の逼迫につながり、他の病気の診療に影響が出るかもしれない。ただし、感染症の拡大という点では、偽陽性が増えてもそれほど問題は生じないと言える。

一方、偽陰性はどうか。偽陰性の受検者は、実際は感染しているにもかかわらず、自宅や医療施設などで隔離されることがない。もし、コロナウイルス感染症のように感染力が強い場合には、偽陰性の人が社会活動を続けることで感染が拡大してしまう恐れがある。つまり、偽陰性は感染症の拡大の点で問題が大きいと考えられる。

上表の感染症の診断の混同行列では、偽陰性の 10 人を減らすために、感染症の診断は積極的に行うべき、ということになるだろう。

◇ 偽陽性を重視するときは適合率、偽陰性を重視するときは再現率が用いられる

このように見ていくと、判断や診断の対象が変われば、偽陽性が問題となる場合や、偽陰性が問題となる場合がある、ということになる。

統計学では、偽陽性を問題にするときには、陽性と判断されたうち実際に陽性だった割合である「適合率」という割合を用いる。先ほどの迷惑メールの判断の例では、40/47＝85.1% といった具合だ。

この適合率が高いほど、迷惑メールの判断はうまくいったということになる。

一方、偽陰性を問題にするときには、実際に陽性だったうち陽性と判断された割合である「再現率」という割合を用いる。先ほどの感染症の診断の例では、20/30＝66.7% となる。

この再現率が高いほど、感染症の診断は(感染症拡大防止の点で)適切だったということになる。

◇ 適合率と再現率を両方用いて評価することが一般的

それでは、他の判断の例ではどうか。気象予報の線状降水帯の発生予報を見てみよう。先ほどまでと同じく、100 回の予報をもとに、混同行列を作る。

線状降水帯予報の混同行列

	線状降水帯発生を予報	線状降水帯発生を予報せず	計
実際は線状降水帯が発生した	10	17	27
実際は線状降水帯が発生しなかった	13	60	73
計	23	77	100

この表の場合は、線状降水帯の予報の有無と実際の発生の有無が一致していたケースが 100 回のうち 70 回(＝10 回＋60 回)あった。

残りの 30 回については、「線状降水帯が発生すると予報したが、実際は発生しなかった」誤りが 13 回。「線状降水帯の発生を予報しなかったが、実際は発生した」誤りが 17 回だった。この結果、適合率は 10/23＝43.5%、再現率は 10/27＝37.0% となっている。

ここで、迷惑メールの判断や感染症の診断とは異なる点が出てくる。線状降水帯の発生予報で悩ましいのは、偽陽性も偽陰性も減らしたい、つまり適合率も再現率も高めたいと考えられる点だ。

偽陽性の場合、被災の可能性のある地域の住民が、予報に従って避難したのに、線状降水帯は発生しないことになる。被災しないという点はよいが、お年寄りや小児などの避難者が、さまざまな肉体的・精神的な負担を被ってしまう点は軽視できない。

一方、偽陰性の場合、予報が出ていないために避難していないなかで、線状降水帯が発生して洪水などの災害が起こり多くの人が被災する恐れがある。

このため、「空振り」の偽陽性と、「見逃し」の偽陰性のどちらも減らしていくことが求められる。

◇ 適合率と再現率を組み合わせた指標「F スコア」を用いて評価することが考えられる

AI の機械学習では、データを分類するモデルの評価尺度として、「F スコア」(※)が用いられる場合がある。まず、「F_1 スコア」として、適合率と再現率の調和平均、つまり逆数の平均の逆数をとる尺度がシンプルなものとして挙げられる。

(※)「F スコア」という名称は、1992 年にコンピュータやコンピュータ科学に関する第 4 回メッセージ理解会議(MUC-4)で紹介されたときに、コンピュータ科学の学者 Van Rijsbergen 氏の著書に記載されていた別の F 関数にちなんで名付けられたとされる。

先ほどの線状降水帯予報の例では、適合率は 43.5%(＝10/23)、再現率は 37.0%(＝10/27)だったので、F_1 スコアは、逆数の平均の逆数をとって、F_1＝1/((23/10＋27/10)/2)＝10/25＝40%となる。

F_1 スコアは、適合率と再現率がともに 100%のとき、最大値 100%をとる。適合率と再現率のいずれかが 0 のとき、最小値 0 となる。

F_1 スコアが大きいということは、適合率と再現率がともに大きいことを意味する。このため、F_1 スコアは、適合率も再現率も高めたいというニーズに合致した尺度となる。

◇ 適合率と再現率のバランスには人間の感覚が必要とされる

F_1 スコアは、適合率と再現率を同程度に重視する尺度である。だが、場合によっては、適合率よりも再現率を重視したいといったこともありうる。

先ほどの線状降水帯の予報の例でいえば、避難者の肉体的・精神的な負担となる偽陽性の問題もさることながら、避難せずに被災してしまう偽陰性の問題のほうが重大だと考えられる。つまり、偽陽性を問題とする適合率よりも、偽陰性を問題とする再現率を重視したい、ということになる。

このように、適合率と再現率のバランスを調整する尺度として、F_β スコアがある。算式で表すと、

次のようになる。

F_βスコア　＝　$\{(1+\beta^2)\times(適合率)\times(再現率)\}$ ／ $\{(\beta^2\times(適合率)+(再現率))\}$

ここで、βは実数とされ、β^2は0以上の値となる。β^2が1のときは、F_βスコアは、F_1スコアと同じになる。β^2が1より小さい場合、F_βスコアは再現率の変化よりも適合率の変化に敏感に反応する。つまり、適合率を重視していることになる。β^2が0のときは、F_βスコアは、適合率に一致する。

一方、β^2が1より大きい場合、F_βスコアは適合率の変化よりも再現率の変化に敏感に反応する。つまり、再現率を重視していることになる。β^2が無限大のときは、F_βスコアは、再現率に一致する。

このようにF_βスコアを使えば、適合率と再現率の重要性をβの設定として盛り込むことができる。すなわち、評価者の感覚に応じて、適合率と再現率のバランスを調整できるわけだ。

通常、AIの機械学習において、βはデータサイエンティストが設定する。両者のバランスをどうとるかは、人間であるデータサイエンティストに委ねられることになる。

複雑なAIの機械学習において、問題の重要性に関する人間の感覚が生かされることとなる。

いま世の中はAIブームの真っ盛りといえるが、まだしばらくは、人間の感覚が必要とされるのかもしれない。そんなことを考えながら、AI開発に関するさまざまなニュースを見てみるのもよいだろう。

(参考文献)

「なっとく！機械学習」Luis G. Serrano著，株式会社クイープ監訳(翔泳社，2022年)

“The truth of the F-measure” Yutaka Sasaki(Teach tutor mater. Vol. 1, no. 5. pp. 1-5., 2008)

“F-score” (Wikipedia)

