第5章 いいデータわるいデータを見極める

ご購入はこちら

正答率をUPさせるための学習データ作りの勘どころ

佐藤 聖

いい人工知能作りは膨大なBADデータとの戦い

● データ品質が悪いとはどういうことか

人工知能時代にはデータの価値が飛躍的に高まります．人工知能では学習するデータをいかにして発見するかが差異化のポイントになります．インターネットから大量のデータを収集するのは容易ですが，そのまま人工知能の学習用データとして利用するには，データ品質にばらつきがあって学習がうまくいかなかったり，同じオリジナル・データから作られたメタデータ（副次的なデータ）が含まれており，学習に利用すると過学習を招き，学習済みモデルの汎用性が失われたりなどの問題が起きる可能性があります．

良いデータを定義することは利用ケースによって異なるため難しいですが，一般的にデータ品質の問題の代表例として**表1**を示します．

例えば画像データや音声データを収集する場合，インターネット上でごく少数のオリジナル・データおよび膨大なメタデータが見つかります．本来であればユニークなオリジナル・データが大量に収集できると，学習するときに汎用性が向上し，認識精度が向上する可能性が高くなり理想的です．

学習用データとしてメタデータが問題になるのは，オリジナル・データと同じ特徴を持つメタデータを大量に学習しても，モデルの汎用性が低くなり，過学習によって特定の特徴しか識別できない特殊な学習モデルになってしまうことです．メタデータは基本的に特徴が同じですが，オリジナル・データのある部分を切り出したモノであったり，改変したモノであったりするため，人工知能で学習できる特徴が少なくなっていることもあります．理想は汎用的に使えて認識精度の高い学習済みモデルが良いはずです．

● たった5%のBADデータが学習をダメにしてしまう

一般的に教師あり学習ではデータセットの5%にミス（無関係なデータの混入，ラベル付けの間違いなど）があると，残りの95%が正しくても，そこから学習して作成されたモデルは誤検知や誤認識などで利用者視点で見ると意外と使えないという印象を与える恐れがあります．たった5%のミスでも学習用データを繰り返し学習したり，複数セットの学習用データを学習したりすると，無視できないような認識精度の低下を招きます．

例えば5%のラベル付けミスのある学習用データ（1000件）があるときに，ディープ・ラーニングの学習の際にクロス・バリデーションによって学習用データを4（800件），テスト・データを1（200件）にした場合，全ての部分を学習しようとすると5回学習を繰り返さなければなりません．学習を繰り返す分正しくラベル付けされた部分の認識精度は向上しますが，間違ってラベル付けされた部分についてもその特徴を学習しているため，5%以上の認識率の低下につながるかもしれません．

● アルゴリズムのチューニングと同じくらい重要

人工知能の認識精度を向上させるためには，アルゴリズムの最適化以前に，学習用データから5%のラベ

表1　データ品質が悪いと人工知能の正答率が下がる
データ品質問題の代表例

データ品質の問題	説　明
データの希薄性	・動画／静止画像が暗い，白飛びなどで不鮮明．夕日やネオンサインの光で変色など ・音量が小さく周辺の騒音で音声が不鮮明など
データの破損	・手書き文字が書かれた紙を静止画像としたものをテキスト・データに変換時に文字以外の記号や絵などに誤変換
無関係なデータ	・お好み焼きの画像セットの中にラーメンやピザ，たこ焼きなど意図しない画像が混ざっている
悪いラベリング	・人間の偏見，データのラベル付け作業（アノテーション）手順によるバイアスや作業ミスなど
予期しないパターンの欠如	・データ収集ルールやデータ定義の欠如による想定外のデータセットが作成されるなど
間違ったパターン	・人工知能のパターン認識において特徴と言えない部分を誤って特徴量として使用するなど