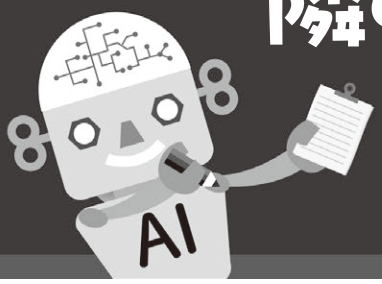
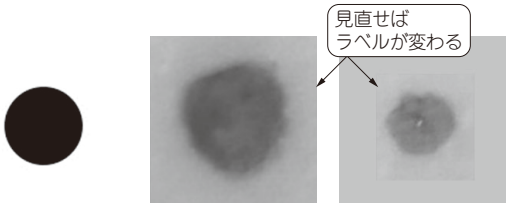
連載

突撃！隣のデータサイエンス



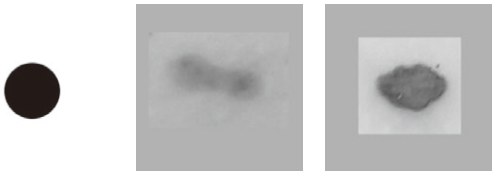
第2回 見直すたびに結果が異なるグレーゾーン・データへの対策

萩野 真一



（a）検査基準　（b）OK　（c）NG

図1　ラベル間違いの例
基準を上回るものにOKがついていたり，基準を下回るものにNGが付いていたりする



（a）基準画像　（b）検査対象1…これはOK？NG？　（c）検査対象2…これはOK？NG？

図2　基準ぎりぎりのデータは判断に迷う
現実の製品異常検査では必ず存在する

前回（第1回，2024年1月号）は有名なデータセットにもラベルに誤りがあることを紹介しました．また，ツールを使ったシミの大きさ分類を体験しました．1回のラベル付け（アノテーション）だけでは不十分で，見直しのサイクルを繰り返すことでラベルの質を上げる必要があることを述べました．

今回は見直し後に残るグレーゾーン・データに対する対策を3つ紹介します．グレーゾーン・データはマルチクラス分類や物体検出，領域分割にもありますが，今回は分かりやすくするために2クラス分類に話を絞ります．

グレーゾーン・データの定義

● 同じ人でも見直すと判定が変わる

人が判断してラベルを付ける場合，**図1**のように後から見直せば誤りが分かるラベル間違いが必ずといってよいほど起こります．ラベル見直しを繰り返すと，**図1**のようなラベル間違いは修正できますが，**図2**のように基準ぎりぎりのデータはNGかOKかを確定することは困難です．

このようなデータはラベル見直しの過程でNG，OKの判断が確定しません．これをグレーゾーン・データと呼びまず．例えば，ラベル見直しサイクルを3回行ったときは，ラベル間違いのデータは3回とも結果が一致しますが，グレーゾーン・データは3回の判断が一致しません．

対策1…領域分割でグレーゾーン・データを減らす

● 目視で判断が難しい場合でも比較できる

前回のツールで使ったシミの大きさを，基準画像を元に分類する問題は，目視でどちらが大きいかを判断するのは困難な場合があります．しかし，シミの面積を計算できれば基準画像との比較は容易になります（**図3**）．

また，領域分割できれば，面積（ピクセル数）を比較することにより基準画像を上回るかどうかを判定できます（**図4**）．ここで，検査対象1［**図2**（b）］のように境界が不明瞭な場合は，どこまでを境界とするかという判断が分かれる要因が残ります（**図5**）．領域分割におけるグレーゾーン・データは，このように境界が不明瞭なものです．

● ラベルの決め方

シミ画像のように2値化アルゴリズムによって2値化できる場合は，しきい値をルールとして定めれば判断の入る余地がなく，明確にラベルを決めることができます．人間が塗りつぶしてラベル付けをする場合は，判断が分かれる要因が残ります．しかし，目視でNG/OKを判断するよりは格段にグレーゾーン・データを減らすことができます．

体験アプリは本連載サポート・ページで確認できます．
https://interface.cqpub.co.jp/202310followup/