9章　教師なし学習のテクニック

1. クラスタリングとは、一般にはインスタンスで似ている群を探索すること、画像のセグメンテーションならカラーチャネルなどの特徴量で似ているピクセル群を探索することである。クラスタリングアルゴリズムには、代表的なものにK平均法やDBSCANがある。
2. データ解析や次元削減、異常検知半教師あり学習のラベル付け、画像セグメンテーションなどに応用されている。
3. まず、クラスタ数ごとの慣性の推移をプロットし、慣性の値の傾きが緩やかになる「ひじ」というポイントに注目し、その付近のクラスタ数にあたりを付けるというやり方がある。また、シルエットスコアを計算し、シルエット図からすべてのクラスタでシルエット係数が十分大きく、インスタンス数のバランスの良いクラスタ数を読み取るという方法もある。
4. ラベル伝播とは、半教師あり学習において、クラスタリングを行って代表データを抽出してラベル付けを人力で行い、それに対して距離的に近いインスタンスにも代表データと同じラベルを与えることである。これによって、ラベル付けのコストを抑えつつ高性能な学習器を訓練することができる。実装には、K平均法などのクラスタリングアルゴリズムを用いればよい。ただし、クラスタ全体に代表データと同じラベルを与えた場合には、境界近くのインスタンスに間違ったラベルが付与される可能性があるため、代表データの近傍のみにラベルを付与するべきである。
5. 大規模なデータセットに対するスケーラビリティが高いクラスタリングアルゴリズムはBIRCHやK平均法。また、高密度の領域を探すクラスタリングアルゴリズムはDBSCANや平均値シフト法である。
6. 例えば、医療分野において、MRI画像からガン細胞と正常細胞を見分けるような分類器を作るような場合である。
7. 異常検知とは、異常なインスタンスとは大きくかけ離れたインスタンスを検知するタスクであり、データセットに異常なインスタンスが入っていてもかまわない。一方で、新規検知は、アルゴリズムが異常なインスタンスの混じっていないデータセットで訓練されていることを前提としている。
8. 混合ガウスモデルとは、パラメータが分からない複数のガウス分布をまぜたものからインスタンスが生成されていると仮定し、一つのガウス分布を一つのクラスタに対応付けてクラスタリングする確率的かつ生成的なモデルである。これは、クラスタリングや異常検知、密度推定などに使える。
9. ベイズ情報量規準や赤池情報量規準が使える。
10. ファイル「code_9.ipynb」参考のこと。おおよそクラスタリングは成功しているように見える。しかし、いくらかのクラスタでは複数人をまとめてしまっている。
11. ファイル「code_9.ipynb」参考のこと。まず、ランダムフォレスト分類器のみで訓練した場合、検証セットの正解率は93.75%であった。K平均法を用いてクラスタリングした後に同じことをすると、正解率は80%であり、PCA次元削減後のデータの特徴量を元のデータセットに追加した場合には、正解率87.5%程度であった。
12. ファイル「code_9.ipynb」参考のこと。何もしない場合はスコアが1000程度である一方、変更を加えた方ではスコアがすべて負の値になった。よって、異常検知ができているといえる。
13. ファイル「code_9.ipynb」参考のこと。再構築誤差の値は、何もしない場合は0.00019程度であるのに対し、変更を加えた方では0.0047であり、およそ30倍も大きい。プロットしたものは、確かに回転などの変更が加えられる前の元の画像を復元しようとしている。