Rを用いた国旗の特徴分類

情報システム工学科 4年 088 山崎隆志

1. 概要

世界各国の特徴データと国旗のデザイン（色や形）との関連性について統計解析ソフト R を用いて解析する。

2. 実験方法

各国の特徴データと国旗のデザインのデータ（クラスラベルデータ）をそれぞれ Web サイトから収集、加工する。

統計解析ソフト R を使用し、特徴データからクラスラベルデータを予測する。「赤色が使われているか?」といった 2 クラスの分類と、「メインカラーは何色か?」といったマルチクラスの分類をする。

予測結果について正答率を求めて結果を評価する。

3. 実験に用いたデータ

各国の特徴データは Web サイト NationMaster(<http://www.nationmaster.com/>) から収集し、7 種類（地域、面積、人口、宗教、言語、農業、GDP）を使用することとした。国旗のデザインのデータは Web サイト UCI(<http://archive.ics.uci.edu/ml/datasets/Flags>) から収集した。

4. 学習器

R の kernlab パッケージに含まれている ksvm 関数を使用した。精度の評価方法は LOOCV(Leave-One-Out-Cross-Veredation) とした。

表 1 : 国旗の色の分類結果

国旗の色	正解の比率	予測正答率%
赤	113 / 149 = 0.75	77.02%
緑	69 / 149 = 0.46	76.35%
青	69 / 149 = 0.46	64.18%
黄	62 / 149 = 0.41	68.91%
黒	43 / 149 = 0.28	76.35%
白	108 / 149 = 0.72	72.97%

5. 実験結果

分類の予測正答率は色の分類、形の分類について行い、2 クラスの分類結果はそれぞれ下の表のようになった。表の項目の「正解の比率」はクラスデータの偏りを表す。

メインカラーを予測するマルチクラスの分類結果は 45.94%であった。

6. まとめ・考察

2 クラスの分類ではデータの偏りによって正しい予測ができないことが多いが、青、緑、黄、黒の分類については正しく予測できたと考える。

国の特徴データから国旗の色を分類することは多少は可能であると考え。しかし、形の分類については難しいと考える。

反省点として、形の分類に関しては異なる学習器を用いて実験すれば異なる結果が得られたと考える。また、分類の作業から作図までしたかったが、今回は形までは予測が難しかったためできなかった。

表 2 : 国旗の形の分類結果

国旗の形	正解の比率	予測正答率%
円	22 / 149 = 0.14	85.13%
十字	15 / 149 = 0.10	91.21%
左上囲	17 / 149 = 0.11	88.51%
太陽・星	61 / 149 = 0.40	43.91%
文字	10 / 149 = 0.06	93.24%
月	10 / 149 = 0.06	93.24%
三角	22 / 149 = 0.14	85.13%
動物	28 / 149 = 0.18	79.05%
縦ライン	29 / 149 = 0.19	80.40%
横ライン	66 / 149 = 0.44	63.51%