平成２１年度自主課題研究概要
～R による生薬・漢方薬の薬効分類～

工学部　情報システム工学科　３年　024　島津　純哉

1.背景

Rの最大の特徴である学習・予測の機能を使って、大きなデータベースの中に現れるパターンを見出し、R の認識関数を使って分類問題を行ってみたいというのが研究の背景である。今回はデータとして、漢方薬の成分データから効能の分類を行った。

2.目的

Rは、商用S言語をオープンソース化したフリーウェアである。インターフェースは、Windows のプロンプトのようなコマンドラインによる対話的操作で、多彩な統計処理を行うことができる。今回の研究で、R の基本的処理を身につけ、漢方薬の成分データベースより効能を分類することを目的とする。また Perl によるデータの抽出の技術も身につける。

3.研究用データ

研究用データは、KEGG　DRUG（http://www.kegg_drug.jp）より、perl を使って必要なデータをまとめた。予測するクラスラベルは

level0（漢方薬剤 or 生薬）

level1（解表剤、解熱剤、健胃剤…29 種類）

level2（辛温発表剤、辛冷発表剤、止血剤…37 種類）

である。

これに各薬剤ごとの成分を 0 と 1 で表した。

4.実験内容

R で分類を行うための識別関数 ksvm により、分類を行った。結果の正答率が以下のようになった。

Level0…98.4%

Level1…38.7%

Level2…34.2%

5.考察

クラスが 2 種類の場合は分類の精度が非常に高いが、クラス数が多い場合は精度が低くなっているのがわかる。R の識別関数では本来 2～10 種程度のクラスの分類には精度が期待できるということがわかった。

6.所感

今回 R を使った統計処理を行ってみて、処理にかけるデータを抽出する行程が重要だと感じた。

また、統計処理には大変多くの時間を要することもわかった。そのため、エラーについても、順次デバッグをしながら対応していくことも学んだ。

7. 総括

このような統計処理は、バイオインフォマティクスの分野や、言語の解析、ビジネスデータ算出などの成果を挙げている。ろう。私も今回 R を使った経験をもとに、さらに幅広い操作や処理を身につけ、自分の強みとして使えるようにしていきたいと考える。