Rを用いた薬の分類

電子情報学類生命情報コース　3年306番武田徳明

1.概要

Rを用いて医薬品の分類を行う。分類は含まれている原子の種類と数を使用する。

2.実験方法

今回Rの分類にはサポートベクターマシンとナイーブベイズという関数を使用する。これに特徴ベクトルである原子の数を学習させ、クラスラベルの薬の種類を予測する。

今回はksvmとnaiveBayesという二つの識別関数を使用し、大まかな分類のlevel0と細かな分類のlevel1のそれぞれで分類を行う。

3.実験に用いたデータ

今回の実験で使用データはKEGG DRUG(http://www.genome.jp/kegg/drug/drug_ja.html)の日本の一般医薬品の分類というページから薬の種類と原子の数を抜き出し、使用した。

4.実験データの例

表１：薬の種類と原子の個数の表

name	level0	level1	C	H	O
アスピリン（JP15）; アセチルサリチル酸	精神神経用薬	かぜ薬（内用）	9	8	4
アスピリンアルミニウム（JP15）	精神神経用薬	かぜ薬（内用）	18	15	9
アセトアミノフェ（JP15）;パラセタモール	精神神経用薬	かぜ薬（内用）	8	9	2

上記のように特徴ベクトルを原子の種類と数、クラスラベルを薬の種類として表を作成した。データのサイズは縦382×横24である。

5.実験結果

表２：分類の正答率

	level0	level1
ksvm	38.50%	16.20%
naiveBayes	3.40%	4.20%

ksvmとnaiveBayesを用いた分類の結果は表２のようになった。

6.考察

ksvmとnaiveBayesでは全く違う分類の仕方をしているようである。全体としてksvmは数の多いクラスラベルに予測結果が偏り、naiveBayesはやや少ないクラスラベルに分類が偏った。

元々これらの学習器は２クラス分類の精度は非常に高いものであるが、多くの種類の分類になるほど精度は落ちていくようである。

さらに今回は特徴ベクトルに原子の数を使用したが、同じクラスラベルの薬でも目立った特徴のようなものは見当たらず、これも正答率を下げる原因になったしまった。いくつかの特徴と組み合わせて予測を行えば正答率が上がるかもしれないが、今回は時間の関係上これ以上のことは出来なかった。