Department of Computer Science & Applied Physics

Identifying E.coli Protein Localization Sites

Overview

Protein localization is the accumulation and concentration of proteins at a specific site in a cell to enable subcellular processes to execute. The *Protein Localization Sites Data Set* from the Institute of Molecular and Cellular Biology at Osaka University consists of 336 instances of 7 features that are highly correlated with the location of proteins in specific areas of a cell.

Class	ID	Description	# Rows
0	cp	Cytoplasm	143
1	im	Inner membrane without signal sequence	77
2	pp	Perisplasm	52
3	imU	Inner membrane, uncleavable signal sequence	35
4	om	Outer membrane	20
5	omL	Outer membrane lipoprotein	5
6	imL	Inner membrane lipoprotein	2
7	imS	Inner membrane, cleavable signal sequence	2

In this practical, we will train a neural network to classify a set of 7 features by predicting the protein localization site that the instance data suggests. Note that there are very few instances of classes 5, 6 and 7 in the training set. This will make it very difficult to train the machine learning model to accurately identify these localization sites.

Exercises

- *Create a new class called* ***EcoliRunner*** and add the declarations for the 2D arrays *data* and *expected* from the file ***ecoli.txt***.

- Use the class *NetworkBuilder* to create a neural network with the following configuration:
  - **Input Layer:** 7
  - **Hidden Layer:** 12 (TanH)
  - **Output Layer:** 8 (TanH)
  - **Alpha:** 0.001
  - **Beta:** 0.95
  - **Epochs:** 10000
  - **Min Error:** 0.0000001
  - **Loss Function:** Sum of Squares

- Create the following data set required to test if the network is fully trained:

```
for (int i = 0; i < data.length; i++) {
   var predicted = (int) net.process(data[i], Output.LABEL_INDEX);
   var actual = Aicme4jUtils.getMaxIndex(expected[i]);
   out.print(predicted + "==" + actual + "\t");
   out.println(actual == predicted ? "[OK]" : "[Error]");
}
```