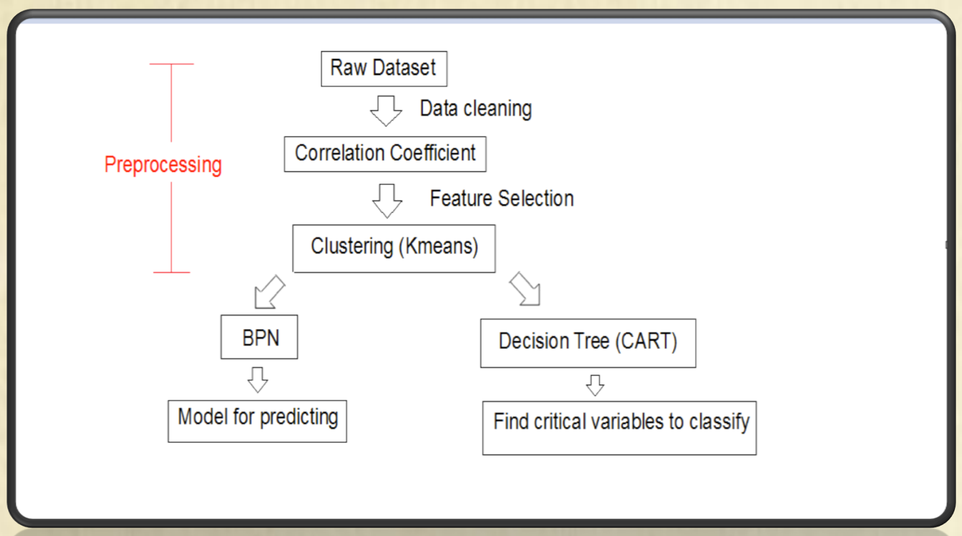
Data mining for bio-chip analysis : a study of breast cancer

Introduction

- Objective :

I find the dataset about patients who had developed distance metastases within 5 years or not on Internet. There are many samples which contain 24481 genes, can we construct a model if you give those genes and the model will tell you relapse or not? And I want to find the critical genes to help us predict.

- Workflow :



- About raw dataset :

High-dimensional biomedical datasets.

The training data contains 78 patient samples, 34 of which are from patients who had developed distance metastases within 5 years (labelled as "relapse"), the rest 44 samples are from patients who remained healthy from the disease after their initial diagnosis for interval of at least 5 years (labelled as "non-relapse").

Correspondingly, there are 12 relapse and 7 non-relapse samples in the testing dataset. The number of genes is 24481 and we extracted values of "Ratio" from original microarray data with the replacement of all "NaN" to 100.0.

Result

- Classification tree analysis :

Purpose : To find the critical variables which are useful when we predict the patient whether relapse or not

As the figure 1 , figure 2 , figure 3 , we can conclude:

1. AL080059:Continuous < -0.1935
   and
   NM_013438:Continuous >= 0.1505
   => output : 1 (relapse) - support : $\frac{6}{96}$ confidence : 100%

2. AL080059:Continuous >= -0.1935
   and
   Contig46934_RC:continuous >=0.34
   => output : 0 (non-relapse) - support : $\frac{4}{96}$ confidence : 75%

3. AL080059:Continuous >= -0.1935
   and
   NM_013438:Continuous < 0.1505
   and
   NM_020244:continuous < -0.546
   => output : 1 (relapse) - support : $\frac{1}{96}$ confidence : 100%

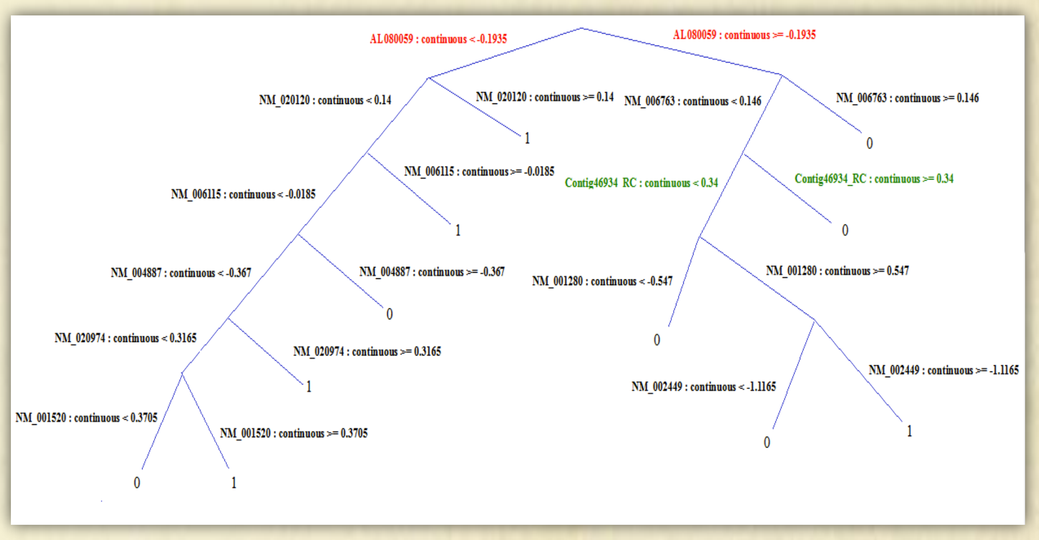


Figure 1 50 variables

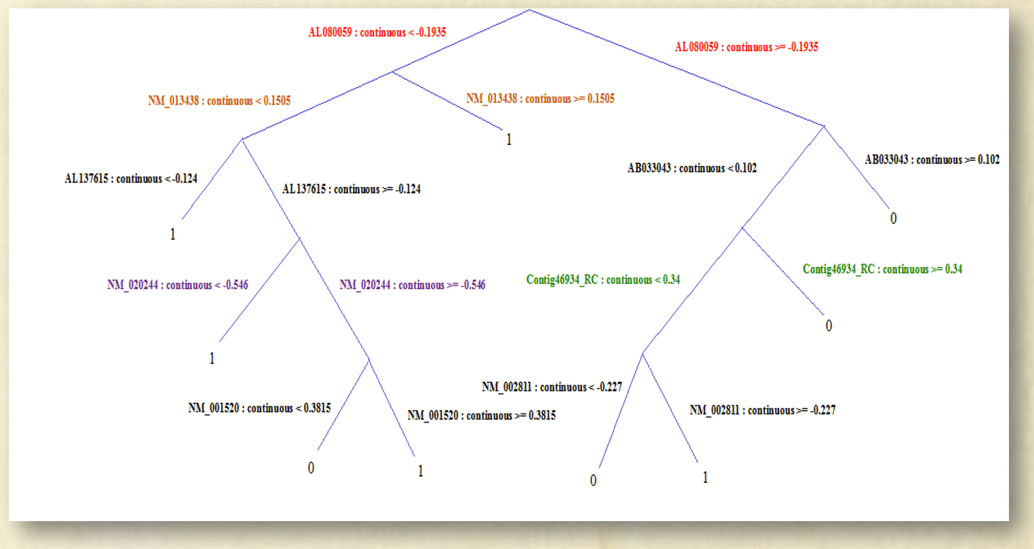


Figure 2 100 variables

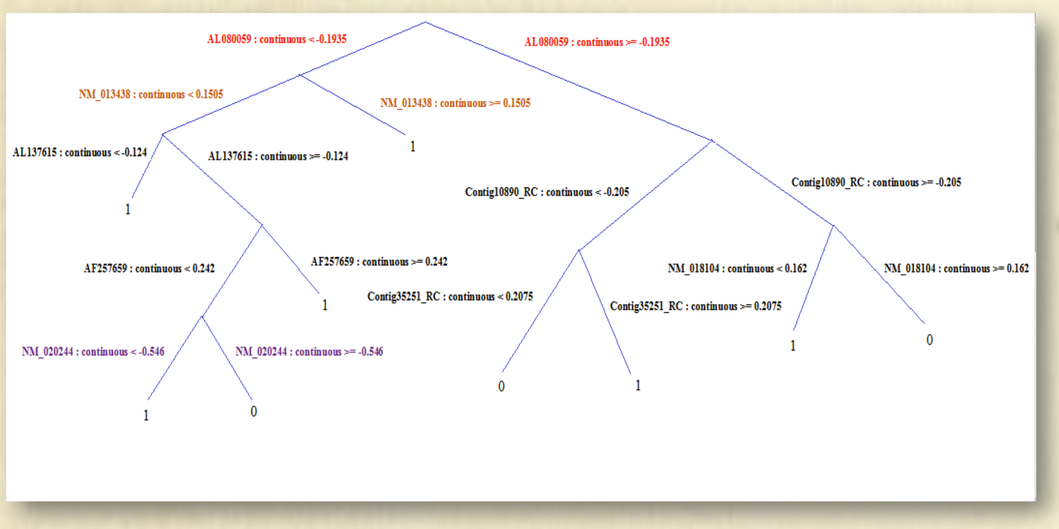
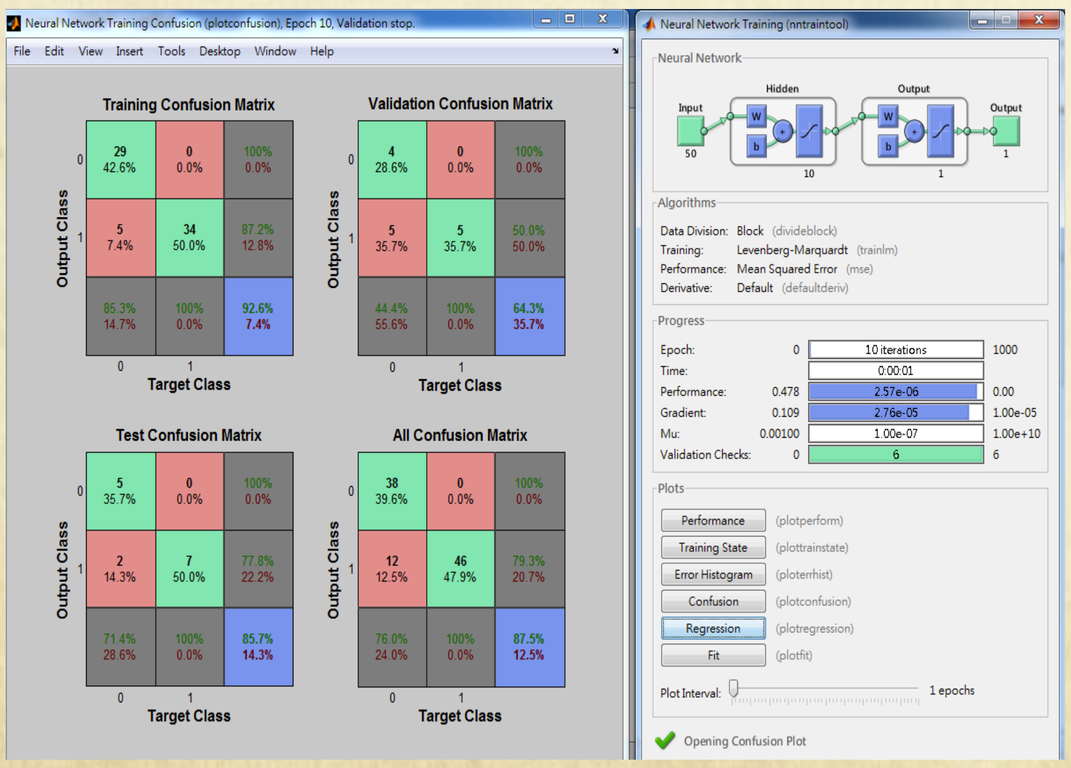
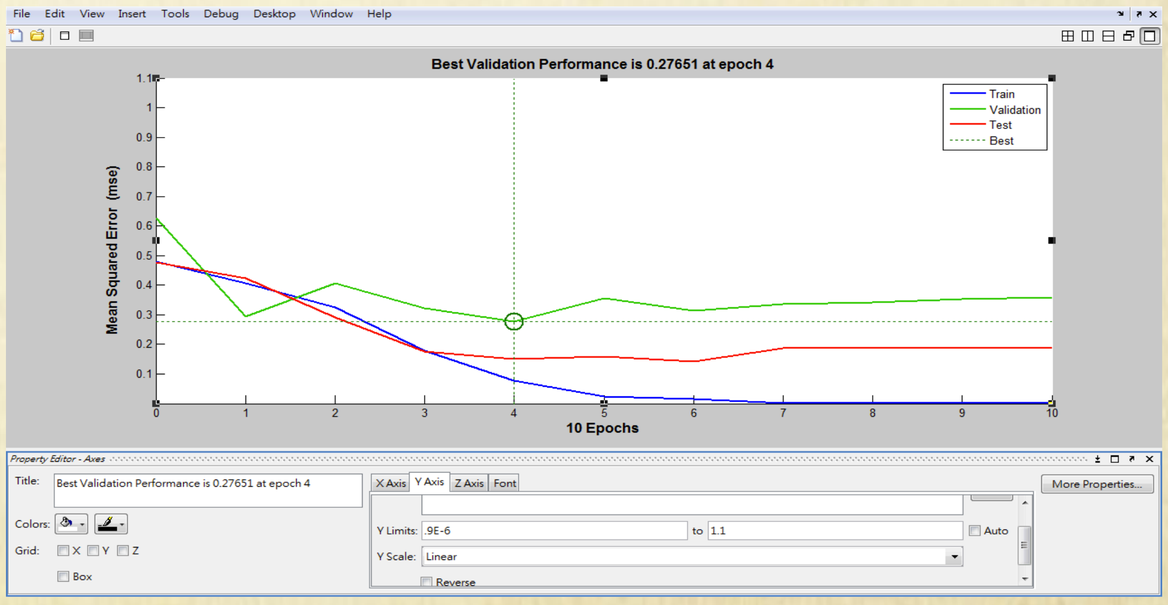


Figure 3
300 variables

- BPN : 50 Variables for example



- 50 Variables : performance



- X axis: label: epochs (number of iterations)
- Y axis: label: Mean Squared Error (MSE)

- Conclusion :

About this project, I construct the BPN model by dataset about breast cancer, if a patient give the those 24481 genes, I can use the model to predict the patient will relapse or not after their initial diagnosis for interval of at least 5 years.

I find the critical genes to determine relapse or not, if you give me the value of those critical genes, I just compare the value and get the prediction result.