# 資料探勘於生物晶片之應用：以乳癌復發為例

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

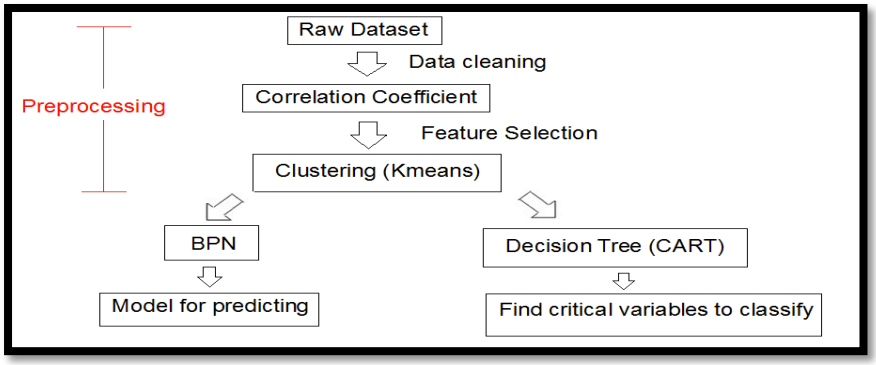
指導教授：李家岩教授
專題成員：顧凱云
開發工具：matlab , C
測試環境：windows 7

## 一.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? Can we find the critical genes to help us predict?

### - Project workflow:




### - About raw dataset:

The training data contains 78 patient samples, 34 of which are from patient 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 data set. 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.

### - Preprocessing:

Use correlation coefficient about input (24188 genes) and output,then extract 1242 genes as the new dataset (correlation coefficient > 0.25 and < -0.25) , the reason for that is simple

, if the input variable has little effect on the output,so I don’ t choose the variable as input to construct the model. (new dataset : 1242 genes as variables ，96 samples => k-means : three version: 50 , 100 , 300 variables version)

# 二. Result:

## - Classification tree analysis:

- Purpose: To find the critical genes 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 : 6/96 confidence : 100% )
2. AL080059:Continuous >= -0.1935 and Contig46934_RC : continuous
   => output : 0(non-relapse) ( support : 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 : 1/96 confidence : 100% )

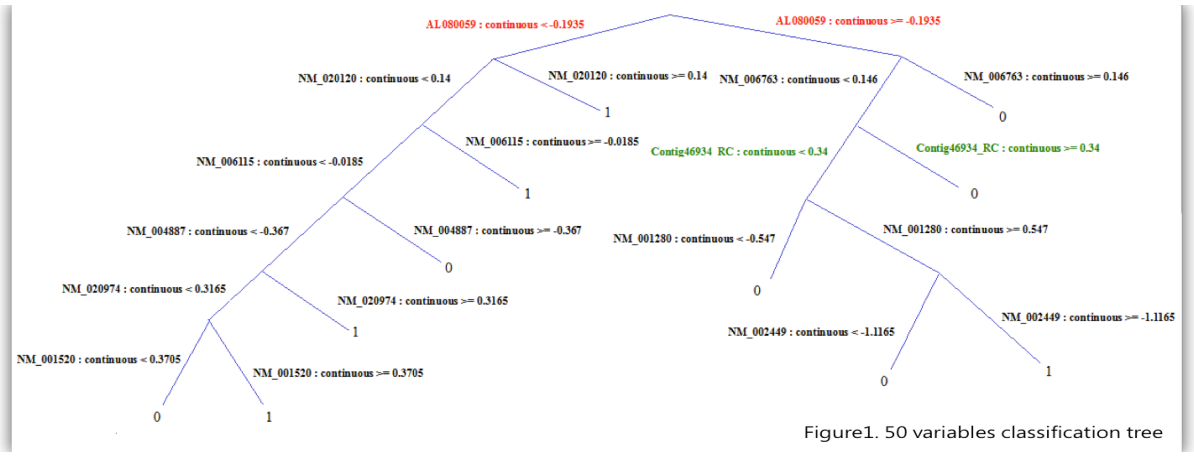


Figure1. 50 variables classification tree

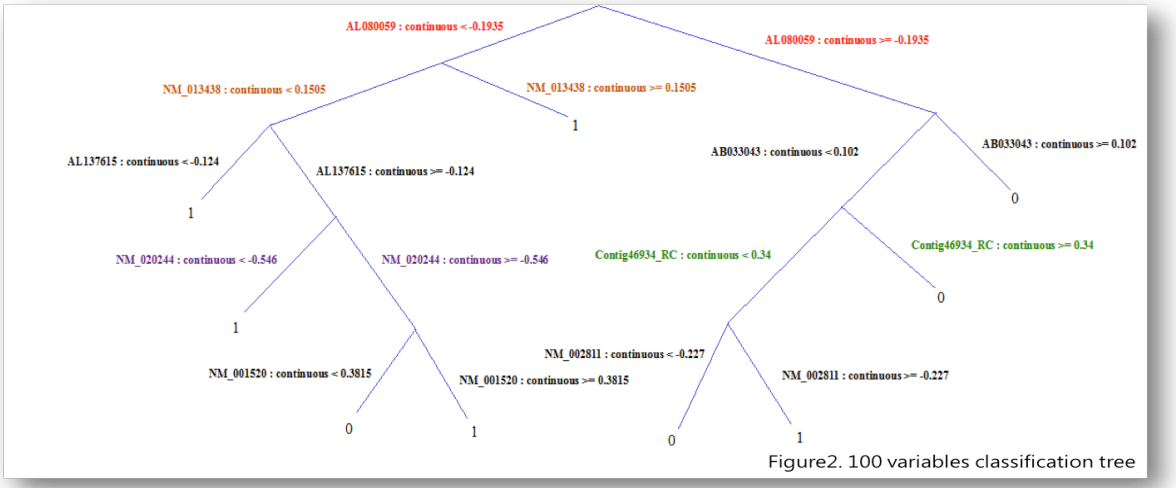


Figure2. 100 variables classification tree

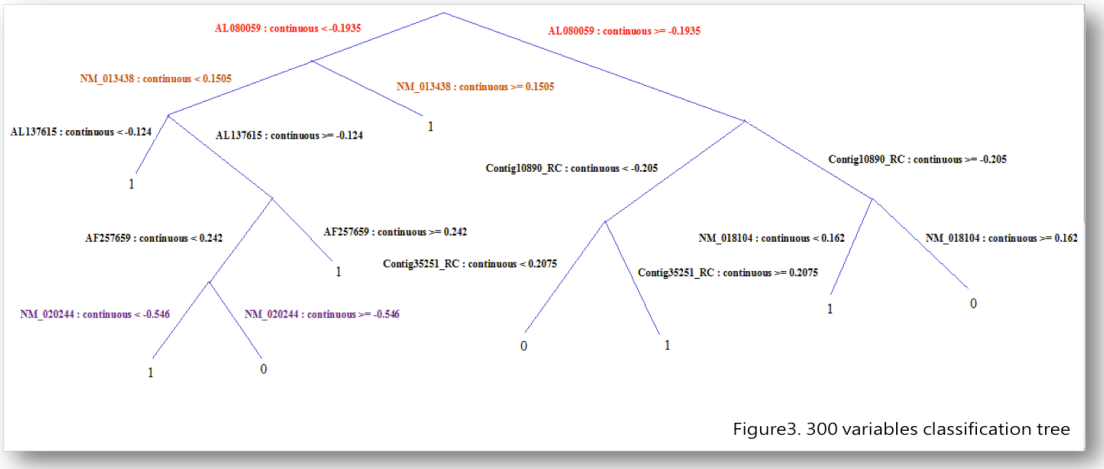
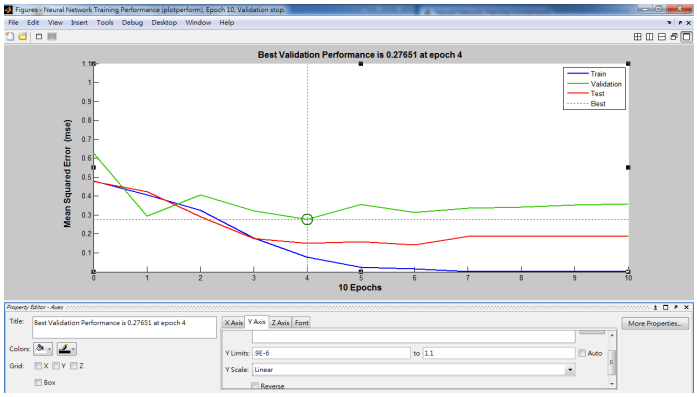



Figure3. 300 variables classification tree

## - BPN :

I use three versions to construct BPN model. Take 50 variables version for example (100 and 300 variables are similar case), 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.