1. Introduction

Study subject RNA is extracted from paired tumor and adjacent normal tissues specimens for gene expression analysis. The dataset contains 54675 genes on sixty paired samples, 60 of which are cancer expressions and the other 60 samples are normal expressions of genes. Genes are Affymetrix expression arrays in each row, all probe sets represented on the GeneChip Human Genome U133 Set are identically replicated on the GeneChip Human Genome U133 Plus 2.0 Array. (Lu et al. 2010)

Purpose:

Results:

2. Method

2.1 Methodology

In this project, several methods are applied to do the classification and clustering. Two primary methods are Nearest Neighbors Classifier and Hierarchical Clustering, which are introduced in the follows.

In pattern recognition, the k-Nearest Neighbors algorithm is a non-parametric method used for classifications and regression. Nearest neighbor classifier classifies objects based on closest training sets in the feature space. The input consists of the k closest training examples in the feature space.

Cluster analysis or clustering is a multivariate analysis to divide data into groups (clusters) that are “similar” to each other but which differ among clusters. The definition of “similar” is varied among algorithms, also the methods of forming clusters vary. In this project, we

used Hierarchical clustering-one of the analysis of clustering to 120 subjects in order to determine how samples in the data set should be grouped into clusters. Hierarchical clustering seeks to build a hierarchy of clusters and it falls into two types: agglomerative-a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. and divisive-a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. (Wikipedia)

2.2 Processing

In classification part, the 60 pairs samples are divided into three subsets in the beginning. Denote nearest neighbor k as odd integer from 3 to 19: {3, 5, …, 19}, then perform the nearest classifier to each k and obtain the number of incorrectly classified subjects. From each of three results, select k with the smallest error rate. Run 5000 random permutations and repeat the steps above to each permutation. Draw the side-by-side boxplot for three results. And the figures are shown in the section 3.

In clustering part, we perform hierarchical clustering for 120 subjects of selected genes based on the values of variances.

3. Result

3.1 Classification

Nearest neighbor classification

15 genes were pre-selected and 60 pairs were divided into three subsets. Nearest neighbor classification (k be the odd integers from 3 to 19) was performed on set 1 as training set and set 1 as test set, set 1 as training set and set 2 as test set and set 1 as training set and set 3 as test set.

Classification error

The numbers of classification error were summarized in table 1 and figure 1.

Table1. Classification errors based on choices of k in three different settings

	K=3	K=5	K=7	K=9	K=11	K=13	K=15	K=17	K=19
Result A	5	2	6	5	8	7	7	7	7
Result B	9	5	7	7	8	7	7	8	8
Result C	8	5	7	7	8	7	7	7	7

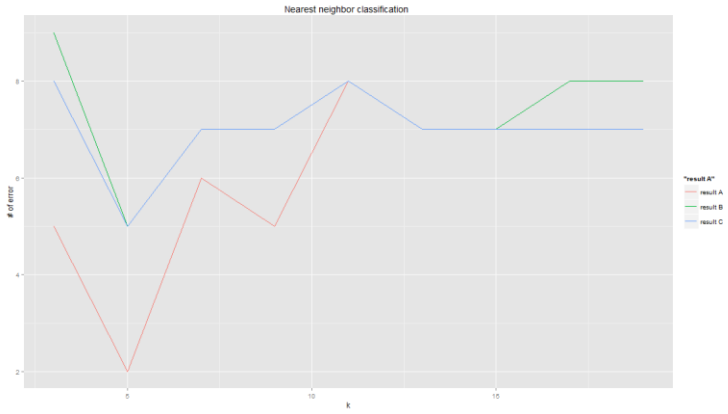

Figure 1. Classification errors based on choices of k in three different settings

(1) Choices of k

K with the smallest error rate was selected. The error rates in three different setting all minimized at k=5 as shown in figure1.

(2) Validation

i. If k is chosen according to result A, then with this k (k=5), the number of errors in result A was 2, error rate was 0.05.

ii. If k is chosen according to result B, then with this k (k=5), the number of errors in result B was 5, error rate was 0.12.

iii. If k is chosen according to result B, then with this k (k=5), the number of errors in result C was 5, error rate was 0.12.

Repetition

5000 random permutations were performed with seed=123. The first, second and last twenty indices were subset into set1, set2 and set3 respectively. Previous steps were performed on the permuted data.

(1) Distribution of classification error was summarized in the following boxplots.

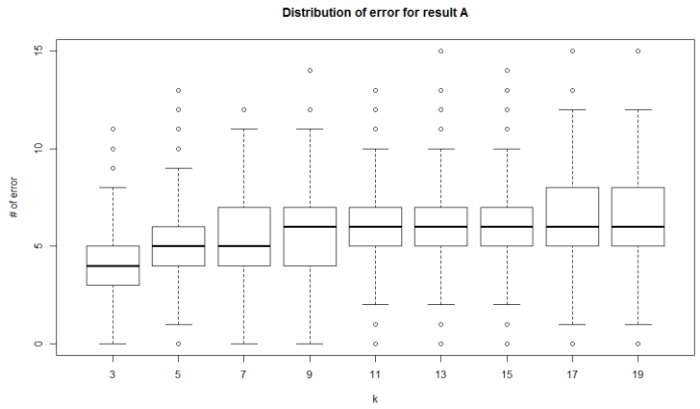

Figure 2. Boxplot of classification error for result A

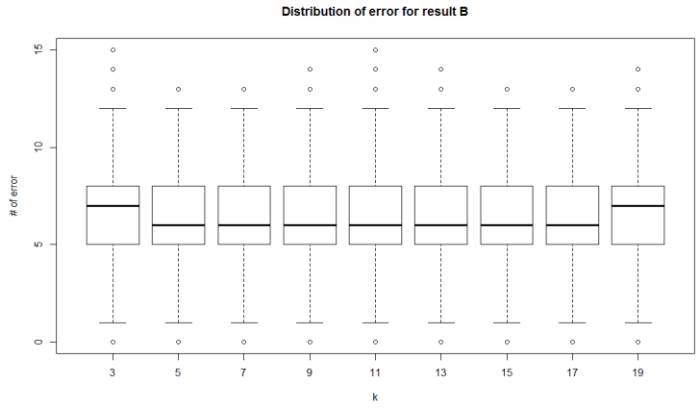

Figure 3. Boxplot of classification error for result B

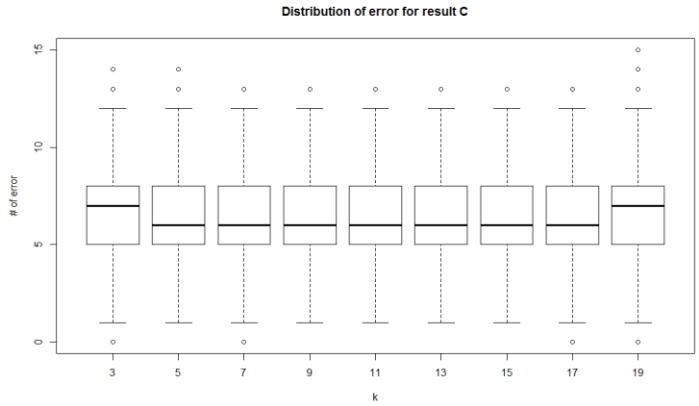

Figure 4. Boxplot of classification error for result C

(2) Distribution of choice of k was summarized in the following boxplot.

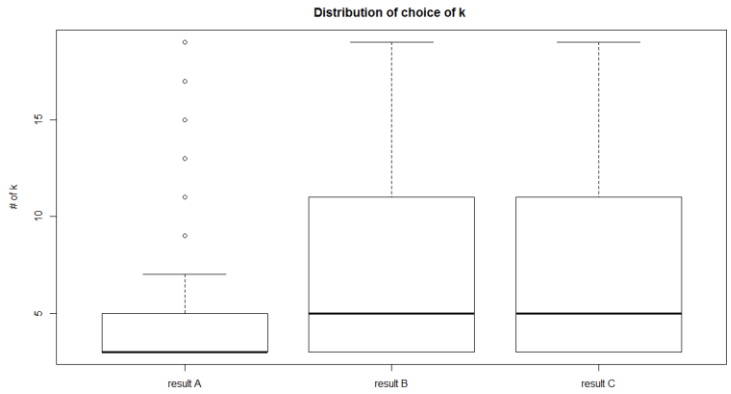

Figure 5. Boxplot for choice of K

(3) Distribution of validation was summarized in the following boxplot.

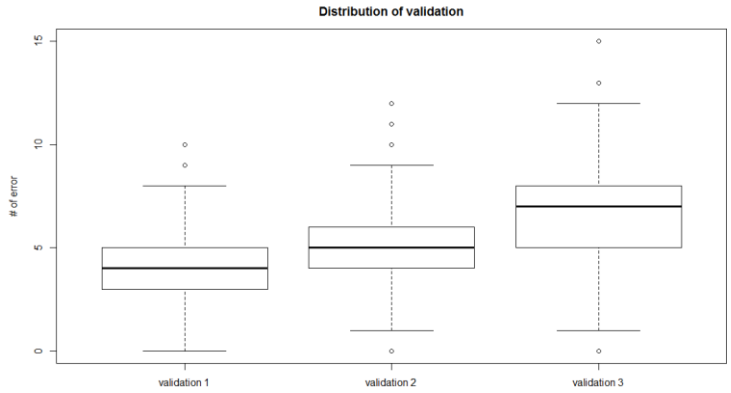

Figure 6. Boxplot for validation

3.2 Clustering

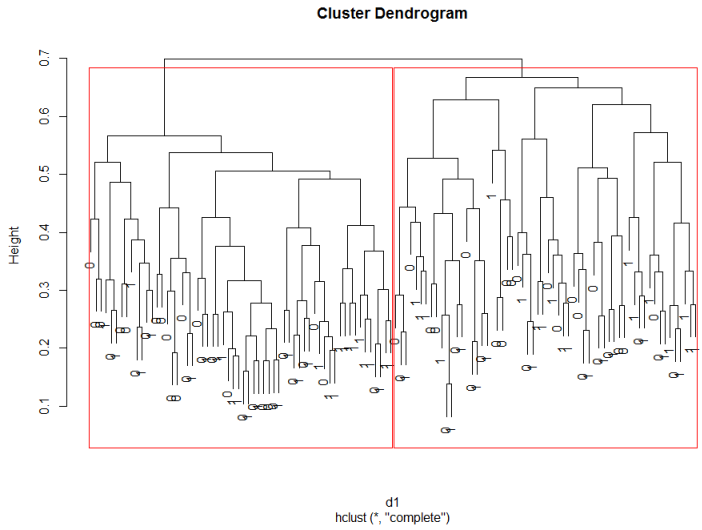

Figure 7. Hierarchical clustering I

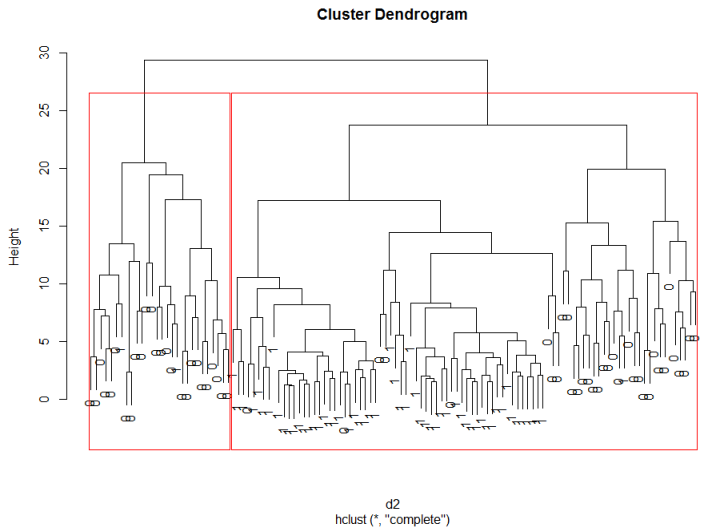

Figure 8. Hierarchical clustering II