Data Mining Project 1 Report

submitted by:Prem Bhusal(U00844184)

In this project we deal with gene expression data for pre-processing and applying feature selection technique. Dataset consist of 62 row and 2000 column as gene attribute and one last column as class(consist of positive and negative).We number each attribute as gene for eg. g1,g2,g3…….g2000 and we perform various tasks which are discussed in detail below.

Task 1:

In task 1 discretization of first K attribute in 3 interval is done. We take one gene attribute at a time and discretize it in 3 interval. As we do equi density binning, we need to put approximately equal number of entries in each bin.For doing this we put each gene attribute in array list and sort it and we find the intervals. We count the entries that falls in each 3 intervals.We compute the variance of each gene

with formula variance = $$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

We represent the output in format as follows.

Gene number: Variance, (-inf,bin1Max],bin1Count ,(binMax1,binMax2],bin2Count,(binMax2, +inf],bin3Count

We store the result in eDensityBin.txt and output looks like this:

G1: Variance :9412169.1777, Bins : [-inf ,5186.0269],21,(5186.0269,7453.4156],21,(7453.4156,+inf],20
G2: Variance :4713963.4423, Bins : [-inf ,3674.192],21,(3674.192,5534.2466],21,(5534.2466,+inf],20
G3: Variance :3252105.1033, Bins : [-inf ,3071.0538],21,(3071.0538,4748.7313],21,(4748.7313,+inf],20
G4: Variance :4010958.5044, Bins : [-inf ,3109.9295],21,(3109.9295,4241.2634],21,(4241.2634,+inf],20
G5: Variance :1811569.0085, Bins : [-inf ,2046.5947],21,(2046.5947,3195.0964],21,(3195.0964,+inf],20

In the second part we map each data entry with bin value we obtain in above task and assign ‘a’ if it falls in first interval , ‘b’ if it falls in second interval and ‘c’ for third interval and we and class value at the end. The out format will be like this:

a,b,c,a,a,b,c,a,b,positive

We store result in file eDensityData.txt and output looks like this:

c,b,b,b,a,b,a,b,c,b,b,a,a,a,b,b,b,a,a,a,negative
c,c,c,b,a,b,b,b,c,c,a,a,c,c,a,c,a,b,b,b,positive
a,c,c,c,a,a,a,a,a,a,a,a,a,a,a,b,a,a,a,a,negative
b,c,c,b,a,a,a,a,b,c,a,a,a,a,a,b,a,a,a,a,positive

a,b,b,b,b,a,a,b,a,b,b,b,b,a,b,a,a,b,a,b,negative
a,a,a,a,a,a,a,b,a,a,a,b,a,b,b,a,a,a,a,a,positive

Task 2:

In task 2 we perform entropy based binning of gene attributes.For each gene attribute we compute the Information Gain considering g1 size-1 splits. We take the split which gives the maximum information gain and divide the bins in two interval. After calculating maximum gain for each gene we rank the gene based on highest to lowest information gain and we take top K gene with information gain ordered in descending order.

Entropy is calculated using formula :

Entropy (E)= -p log p -n log n ------(1) : (log is base 2)

Suppose v splits the each column in two split s1 and s2

So the information of split will be

IS(s1,s2) = s1/s* Entropy(s1) + s2/s*Entropy(s2)

Information Gain = Entropy(s) - IS(s1,s2).

We represent the output in following format:

Gene number: Info Gain, (- , maxSplit], count, (maxSplit,+], count

We store the result in entropyBins.txt and looks like this:

g1915: Info Gain: 0.1985; Bins: (-, 78.3226] , 22; (78.3226, +] ,40
g1362: Info Gain: 0.1977; Bins: (-, 124.4888] , 36; (124.4888, +] ,26
g339: Info Gain: 0.1884; Bins: (-, 135.16] , 7; (135.16, +] ,55
g1392: Info Gain: 0.1838; Bins: (-, 157.8906] , 21; (157.8906, +] ,41
g1392: Info Gain: 0.1838; Bins: (-, 157.8906] , 21; (157.8906, +] ,41

For the second part we map the original data with partition obtained from above implementation . We print data falling into first interval as ‘a’ and that falling in second interval as ‘b’ such as
a,b,b,a,b,a,a,b,positive

Output is stored in entropyData.txt and looks like this below:

b, b, b, b, b, b, b, b, a, b, b, b, b, b, a, a, a, a, b, a, negative
a, b, b, b, b, b, a, a, a, b, b, b, b, b, b, b, b, a, a, a, positive

a, b, b, b, b, b, a, a, a, a, a, a, a, b, a, a, a, a, a, a, negative
a, b, b, b, b, b, b, b, a, b, a, a, a, a, a, a, a, a, a, a, positive
a, a, b, a, a, b, a, a, a, a, b, b, b, b, a, a, a, a, b, a, negative
a, a, a, a, a, a, a, a, b, a, a, b, b, b, a, a, a, a, a, a, positive
a, b, b, b, b, a, a, a, a, b, b, b, b, b, a, a, a, a, b, a, negative

Task 3:

For the third task we need to compute correlation coefficient among the gene and we sort the pair of gene based on highest to lowest correlation coefficient value.

Formula used for calculation of correlation is :

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$ where Cov(A,B) is covariance .

Covariance is calculated as

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

Output of result will be like as follows

(geneA, geneB) : correlation coefficient

The result will be stored in corellatedGenes.txt and looks like this:

(g513, g1808) :0.8795
(g1227, g1915) :0.7852
(g1162, g1088) :0.7739
(g1231, g510) :0.7685
(g1231, g510) :0.7685
(g1162, g1915) :0.7588