CSE514 Spring 2018 – Datamining Homework 2

Assigned date: 2/26/2018 **Due date: 3/20/2018**, before class starts

Tips: Abide to the collaboration policy – see course administrative info or talk to the instructor for details. You may write your own code or use WEKA or other software tools of your choice.

**Preamble –**

Feature selection is an important topic in DM. Instead of giving lectures on the topic, I like to let you get familiar with this topic by working on the issue using two different approaches, which are considered here.

In a disease study, we typically contrast and compare a group of disease subjects against a group of normal controls in order to find a set of genes that are correlated to the disease phenotypes. So for this type of problems, we are dealing with two class problems.

Here we consider the problem of finding critical genes whose aberrant regulation (or expression) may lead to Alzheimer’s disease, the most devastating dementia in the aging population.

**AD dataset –** Download AD.zip file from the Resources page (in the General Resources page) on the course piazza website. This is a real dataset of more than 8000 genes of 176 patients of Alzheimer’s disease (in text file case.gex) and 188 age-matched normal people, or controls (in text file ctrl.gex). You may load them into MS Excel to have a better view. If you want know how these data were collected and used in a disease study, see the original paper: <http://www.sciencedirect.com/science/article/pii/S0002929709001086>.

Due to possible memory issue and to make your hw (project) manageable, we will use the **first 800 genes** for each of the two datasets for all 3 problems in this HW.

**P1:** Feature selection– Random Forest (RF) based (20 pts)

Using the control data as the positive data and patient data as negative data for training (no test data) to build the following RFs, each of which has 500 trees. Compute the frequency of each gene appearing in each of the following RFs.

1. Fix the number of features (genes) at 70% of the total number of features, vary the number of instances from 50% to 90%, with an increment of 5%;
2. Fix the number of instances at 70% of the total number of instances, vary the number of features (genes) from 50% to 90%, with an increment of 5%;

Consider and answer the following questions:

1. What does the frequency of a gene in a RF tell us about this gene for AD?
2. Or, how do we compare two genes based on their frequencies?
3. Now the feature selection: Not all 8,000+ (or 800 in this problem) genes are important or useful. If you can only use a small number of genes, say 100, which ones do you use?
4. Not required, but if you do it correctly or in a meaningful way, you get 5 bonus points) Are the orderings of the genes using the RF method stable? What factors affect the stability? How can we quantify the stability?

**P2:** Feature selection – SVM (20 pts)

Recall that the maximal margin classifier of SVM is f(x) = (wTv+b), where vector w and scalar b are the parameters of the SVM model learned from the training data (in this case, the data are the data for the AD cases and normal controls combined). We now use the following kernels to build SVM models for AD

1. Linear kernel: k(x, x’) = (xTx’), whose model is simply f(x) = (wTx+b);
2. Homogeneous quadratic kernel: k(x, x’) = (xTx’)2.
3. Inhomogeneous quadratic kernel: k(x, x’) = (xTx’+1)2.

Consider and answer the following questions -

1. Given an SVM model f(x) = (wTv+b), i.e., w and b are given, what do the values in vector w tell us about individual dimensions (features)? Hint: how can we compare the importance of two features?
2. With the linear kernel, how should we select the top features? Your answer gives rise to a feature selection method. Now choose the top 100 genes using your method and compare them with the result from P2(4) above.
3. Which of the two quadratic kernels is better for feature selection and why?
4. With the two quadratic kernels, can we find correlated features (genes)? Why?

**P3**: One step further beyond feature selection – Association Rule (AR) mining (20pt)

**The problem**: We are going to use the AR method to find possible gene regulation in Alzheimer’s disease using the AD data. The purpose here is to show you how a difficult biology problem can be approached with a datamining technique.

In the first two problems, you performed feature selection on the AD data using random forest and SVM. Selection of features is just one way to get rid of features that are not important to the problem at hand. Another way is to explicitly identify and remove redundant features. Now, using the results of ARs that you learned from the data, develop a method

AR can also be used to identify redundant features.

**Data preprocessing**: We first need to generate basket (transaction) datasets from the gene expression data we have. Since we have datasets for AD cases and normal controls, we can have two sets of basket datasets, respectively, one for cases (in case.gex) and the other for controls (in ctrl.gex). In one dataset (for AD or controls), each gene can be viewed as an item and each person can be viewed as a transaction. We now consider or determine whether a transaction (person) has a particular item (gene or gene expressed) as follows. First, a person having a particular gene means that she has that gene expressed to have abundance above a certain level. Since the genes that are not highly expressed may not be interesting, we will discard them from our baskets. Therefore, for each feature (gene), we will only consider the x% people whose expressions of that gene are among the most abundant. For this problem, we use x=10%. This means that we turn the gene expression matrix of the given data into a new 1/0 matrix, where (i, j) = 1 if the i-th person has the expression of the j-th gene (item) among the top 10% highest expression levels, or 0 otherwise. Take 10% and the 187 individuals in the control dataset as an example. In this case, we should expect to see 18 entries (or people) have 1 for a particular gene or 0 for the rest.

**Tip**: To help you make sure your preprocessing is correct, we provide the converted data for the first 10 genes for the control data and the total expressed genes (entries with 1 in the matrix) for each person in the file ctrl\_10percent\_confirm.xlsx in the AD data package. (Note, this 10% dataset has the results for the first 1800 genes, so you only need to look into the first 800 genes.)

**What to do**:

1. Given the two basket datasets created above, find all frequent itemsets and interesting ARs with the minimal support from 6% to 10% (with an increment of 1%) and the minimal confidence from 50% to 90% (with an increment of 10%). Draw 3D plots, for AD and controls separately, showing the total number of frequent itemsets and ARs discovered (in separate figures) as a function of the min support and min confidence.

Tip: make sure you turn the 1/0 matrices into the right basket data format for the software you use.

1. Sort the frequent itemsets with supports greater than their corresponding min support, from the highest frequency to the lowest, and report the top 50 most frequent itemsets. Also report all the ARs whose confidences are greater than the min confidence level of 80%.
2. Answer the following question: What information does a discovered AR give us? Or, What is the meaning of an AR here?
3. Design a strategy to identify independent features (genes) and features (genes) that are dependent on the independent features using the result from AR mining.

Tip: You only need to describe the main idea and rationale of your strategy, and a pseudo code may help. The key should be your rationale.

1. Apply your strategy to the results of 2) above (i.e., ARs (confidence > 80%) from the top 50 most frequent itemsets with support > 5%) and list the independent features and their associated redundant features (genes).