Data Mining with Weka
Heart Disease Dataset

1 Problem Description

The dataset used in this exercise is the **heart disease** dataset available in `heart-c.arff` obtained from the **UCI repository**[1]. This dataset describes risk factors for heart disease. The attribute `num` represents the (binary) class attribute: class `<50` means no disease; class `>50_1` indicates increased level of heart disease.

The main aim of this exercise is to predict heart disease from the other attributes in the dataset. Obviously, this is a classification problem. The software to be used is Weka 3.6 . However, feel free to try any ideas you may have to tackle the problem with any other software.

The description of this exercise is stepwise. Therefore, I hope you can get a better understanding of the various aspects and questions involved in the **KDD** process.

1.1 Data Understanding

The first step in approaching the problem is to get acquainted with the data. Answering the following questions will help you to better understand the data.

The data file `heart-c.arff` contains some information about the data stored in it. You can open it with a text editor.

Load the data file in Weka.

1. For each attribute find the following information.
   (a) The attribute type, e.g. nominal, ordinal, numeric.
   (b) Percentage of missing values in the data.
   (c) Max, min, mean, standard deviation.
   (d) Are there any records that have a value for the attribute that no other record has?
   (e) Study the histogram at the lower right and informally describe how the attribute seems to influence the risk for heart disease. What does

[1] A link to UCI data repository is available from the course web page.

it mean the pop-up messages that appear when dragging the mouse over the graphic?

(f) Are there any outliers for the attribute under consideration?

i. Investigate the possibility of using the Weka filter `InterquartileRange` to detect outliers[2].

2. Switch to the `Visualize` tab on the upper part of the screen to visualize 2D-scatter plots for each pair of attributes.

(a) Which attributes seem to be the most/least linked to heart disease? Summarize in a table your findings concerning the predictive value of each attribute.

(b) Does any pair of attributes seem to be correlated?

3. Investigate also possible multivariate associations of attributes with the class attribute, i.e. study scatter plots of two attributes X and Y and try to identify possible "dense" heart disease areas (if any).

(a) If you find "dense" heart disease areas in any scatter plot then quantify the heart disease rate in these areas with respect to the entire data set.

1.2 Data Preprocessing

The second step is to preprocess the data such that the transformed data is in a more suitable form for the mining algorithms.

1. Attribute selection.
Investigate the possibility of using the Weka filter `AttributeSelection` for selecting a subset of attributes with good predicting capability. Then, describe briefly the filter(s) you used and compare the results you obtained with the conclusions you obtained in the previous section. Save the dataset with the selected attributes in the file `heart-c1.arff`.

2. Handling missing values.
Consider the following methods for handling missing values and investigate each possibility within Weka. Note that, as rule of thumb, if an attribute has more than 5% missing values then the records should not be deleted and it is advisable to impute values where data is missing, using a suitable method.

(a) Replace the missing values by the attribute mean, if the attribute is numeric. Otherwise, replace missing values by attribute mode (if the attribute is categorical). Save the dataset you obtained without missing values in the file `heart-c2.arff`.

[2]See also the information about box plots available from course web page.

(b) Investigate the possibility of using (linear) regression to estimate the missing values for each attribute. Save the dataset you obtained without missing values in the file `heart-c3.arff`.

3. Eliminating outliers.
   Eliminate the outlier records and save the dataset you obtained without outliers in the file `heart-c34.arff`.

1.3 Mining the Data

The third step is to use some classifier algorithms available in Weka to discover hidden patterns in the data.

You should repeat the steps described below for each of the datasets you created during preprocessing, besides using also the original dataset (if possible).

1. Start with `OneR` classifier.
   (a) What can you conclude? Compare your conclusions with your previous conclusions obtained in section 1.1.
   (b) Compare the accuracy of the classifier on the training set with the accuracy estimation obtained through 10 fold-cross validation. How do you explain the difference (if any)?
2. Use `JRip` classifier, i.e. the Weka version of the rule classifier `RIPPER`.
   (a) Build a classifier with and without rule pruning. Which one is preferable? Motivate your answer.
   (b) Describe the patterns you obtained and compare with your previous conclusions.
3. Use `J48` classifier, i.e. the Weka version of the decision tree classifier `C4.5`.
   (a) Investigate the use of different `J48`'s parameters such as pruning and minimum number of records in the leaves.
   (b) Describe the patterns you obtained and compare with your previous conclusions.

1.4 Clustering Tendency

Investigate whether there is a clustering tendency in the dataset. You may start by clustering the data with `SimpleKMeans` algorithm, for some $2 \leq k \leq 10$.

1. Do not use the class attribute, `num`, for clustering.
2. Find a suitable value for k, i.e. the number of clusters you are going to build. Justify your choice of k.
3. Use class to cluster evaluation and make sure that standard deviations are also computed for numerical attributes.

4. Study the numerical measures displayed by Weka for each cluster. What can you conclude?

5. Use `Visualize cluster assignments` and try to discover a description for each cluster.

6. Investigate the possibility of building a classifier for finding rules describing the clusters. Compare the results with your previous conclusions.

7. Investigate the possibility of using the cluster information to build a classifier for `num`. Compare your results with what you obtained in section 1.3. Do you get a better classifier?

1.5 Predicting Performance

In the previous step you have built several models. Finally, you need to compare the different models and describe your final conclusions.

1. Weka outputs several performance measures. Choose some of the performance measures and motivate your choice.

2. Summarize in a table the performance measures for each classifier and each dataset.

3. What can you conclude?

1.6 Conclusions

Describe your final conclusions and indicate which risk factors for heart disease have you found in the data.