Project task for ML410C “Projects in health informatics - Project and information management” – data mining

Panagiotis Papapetrou

Data

For this project you will use the UCI Breast Cancer dataset available here:
http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

The original dataset contains 569 examples with 32 attributes. We are going to study only a part of this dataset that is the *Winscosin version* and is contained in files *breast-cancer-wisconsin.data* and *breast-cancer-wisconsin.names.* The input file should be *breast-cancer-wisconsin.data* and it contains the following attributes:

1) ID number
2) Diagnosis (4 = malignant, 2 = benign)
Ten real-valued features are computed for each cell nucleus. Note that these features have already been discretized to 1-10:

1. Clump Thickness	1 - 10
2. Uniformity of Cell Size	1 - 10
3. Uniformity of Cell Shape	1 - 10
4. Marginal Adhesion	1 - 10
5. Single Epithelial Cell Size	1 - 10
6. Bare Nuclei	1 - 10
7. Bland Chromatin	1 - 10
8. Normal Nucleoli	1 - 10
9. Mitoses	1 — 10

Important note: You should be looking at the file called “**breast-cancer-wisconsin.data**”. The patient ID appears first, then the nine features above, and finally the diagnosis. All features values, IDs, and Diagnosis are separated by a comma. More information about the data can also be found in file “**breast-cancer-wisconsin.names**”. All other data files should be ignored!

Objectives

The goal of this project task is to become familiar with a data mining tool of your choice (e.g., Weka) and use it for building two data mining models for classification.

Task 1: Decision tree
Build a decision tree classifier on the Diagnosis attribute (malignant, benign). You are free to decide which attributes to use for building the tree. Note that not all attributes may be necessary. Show the graphical representation of the decision tree that you have obtained and discuss which attributes you have used and why.

Task 2: Evaluation A
Describe the approach you used for training and testing the classifier. Report the accuracy, precision, recall, and F-measure of your classifier. Also, show the confusion matrix.

Task 3: K-NN classifier

Build a 1-NN classifier on the Diagnosis attribute (malignant, benign). Note that you may again choose a subset of attributes for the classifier. Discuss what distance measure you have used for the classifier and what was the rationale for this selection.

Task 4: Evaluation B
Describe the approach you used for training and testing the 1-NN classifier. Report the accuracy, precision, recall, and F-measure of your classifier. Also, show the confusion matrix.

Task 5: Model comparison
Compare the two models you have created and discuss which one you would recommend to domain experts for that dataset. Note that you should describe the approach you used to compare the models.

Tools
Any tool is acceptable. Some suggestions include:

- Weka
- Rule Discovery Tool
- R
- Matlab

Deliverable
A report of a maximum of 5 pages that contains:

- One section with a short description of the contributions of each participant
- Five sections (one for each task) where you describe your approach and results as well as any difficulties or comments related to each task