Evaluation of Q5

Authors: Ranjani Ramakrishnan, Ted Laderas

The outline of the document is as follows:

1.0 Overview

The Matlab implementation of Q5 was used for evaluation purposes. Scripts were obtained from http://www.cs.dartmouth.edu/~donaldlab/Software/. The analysis pipeline consists of the following steps: data is first split into training and test sets. The dimensionality of the data is reduced using principal components analysis (PCA). The PCA step is followed by Linear Discriminant Analysis (LDA) for minimizing the within-class scatter and maximizing the between-class scatter on the training set. The output of the LDA step is a linear discriminant that is used to classify the test set samples. Because the classification of the test set is known, the performance of that particular discriminant can be evaluated (described in section 3). A new test/training split is then picked and the entire process repeated. A number of iterations of the process are completed and the average behavior of the discriminants is calculated and reported. Q5 has been tested on low-resolution SELDI-TOF data to discriminate between serum samples from cancer patients (Ovarian Cancer and Prostate Cancer samples were used) and normal cancer patients.

2.0 Installation

Installation of Matlab was a prerequisite for executing the Q5 scripts. The process of running the Q5 scripts was straightforward and was described in the README file distributed with the code.

3.0 Data Analysis

Current inputs to Q5 are in the form of four files, containing the mass-to-charge ratios (m/z) and relative intensities (RI) for the cancer and normal spectra respectively. The tool is run from the command line for Matlab and the inputs are specified as parameters passed to the different functions, called in sequential order. Because the evaluated version of Q5 is a proof-of-concept version of the software, data for training and testing are input into the program at the same time. The current outputs for Q5 are graphs for the different threshold levels that give the sensitivity and specificity of the classification, the percentage of samples classified, percentage of samples correctly classified and the positive predictive value (PPV).

4.0 Usability Issues of the Matlab Version of Q5

i) Q5 is currently run from the command line by calling each function sequentially. Implementation of the GUI, as proposed, would make this more user friendly.

ii) Currently, the input to Q5 is in the form of four data matrices. Using data directly from the LIMS or in a post-LIMS scenario might require some massaging to get into the form required by Q5. In the R version, there should be a clear decoupling of the inputs to Q5, namely the training spectra and the unclassified spectra.

iii) The outputs of Q5, currently in the form of graphs, would need to be modified. At a minimum, the classification of the test samples into one of the categories would be the output. Depending on whether the implementation is the standalone version (Use Case 3.2, Requirements document) or the LIMS compatible version (Use Case 3.2, Requirements document), the output would be directed to a user or logged into the database.

iv) Currently, the internal objects used by Q5 are not transparent to the user. Grid enabling the implementation would require rethinking regarding which of the objects will be visible to the grid and the format in which they will be presented.

v) Lastly, the validation of Q5 has been carried out using (comparatively) low resolution SELDI-TOF data. Validation of the implementation for high resolution MALDI-TOF data is needed, in addition to testing.

5.0 Relevant Documents

Draft of Use Case Document for Q5