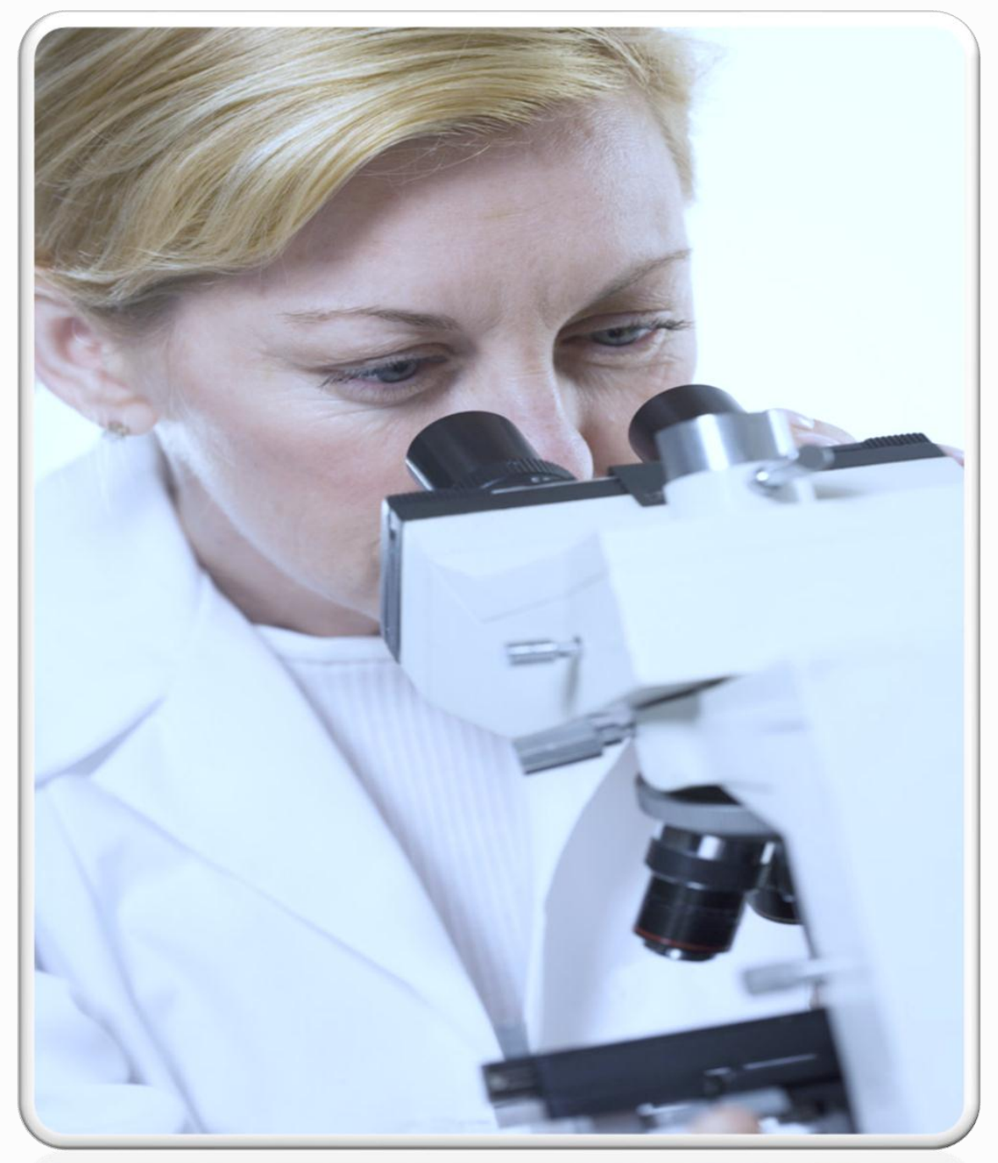
Classifying Spectroscopy Data



Author: Sheheryar Khan **Supervisor: Plamen Angelov**
2nd Supervisor: Dr F. Martin (Biology) Assistant: J. Trevisan
Lancaster University – Communication System Department – InfoLab21

Abstract

Despite huge progress in the understanding , prevention and treatment, the oncology disease remains leading mortality rate. Infrared spectroscopy is a potentially new method for cancer diagnosis, due to sensitivity of the technique to alterations in cellular biochemistry which accompany disease stages.

This thesis examines the use of intelligent computational methods in the development of predictive models for medical classification based on FTIR spectroscopy data. In particular the evolving fuzzy rule based classifier (eClass), Distance To Weighted Mean classifier, 1-R/Class Fuzzy classifier and Linear Discriminant Analysis have been our focus. We investigated feature selection using eClass that enhances the predictive performance of classifiers and most importantly helps identification of different biomarkers that are more useful for biological point of view. The feature transformation Principal component analysis along with Linear discriminate analysis results in highest system accuracy with Minimum distance classifier and fuzzy classifier.

Introduction

The data processing has three stages: pre-processing, feature extraction and classification. This poster explains selected methods. For detailed explanation of the full project, please contact the author for a PDF version of the dissertation.

FTIR Spectroscopy

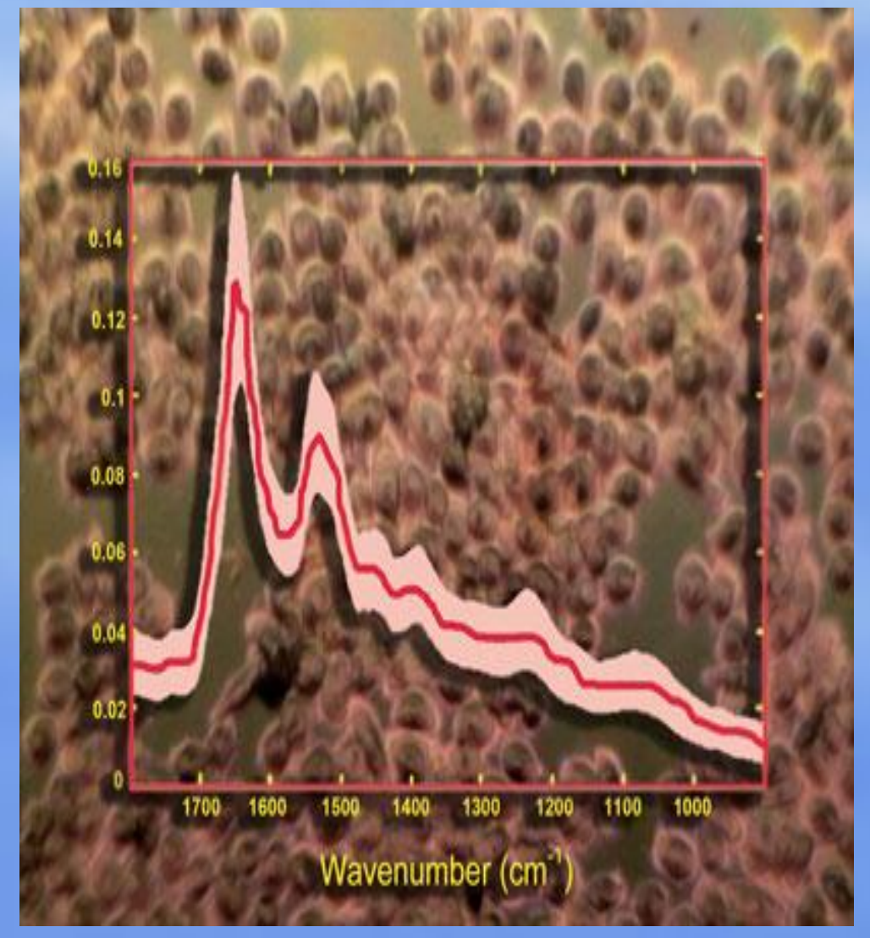
Primary screening tools

- ❖ Morphology
- ❖ Subjective interpretation
- ❖ Require skilled pathologist

Promising alternative - FTIR Spectroscopy

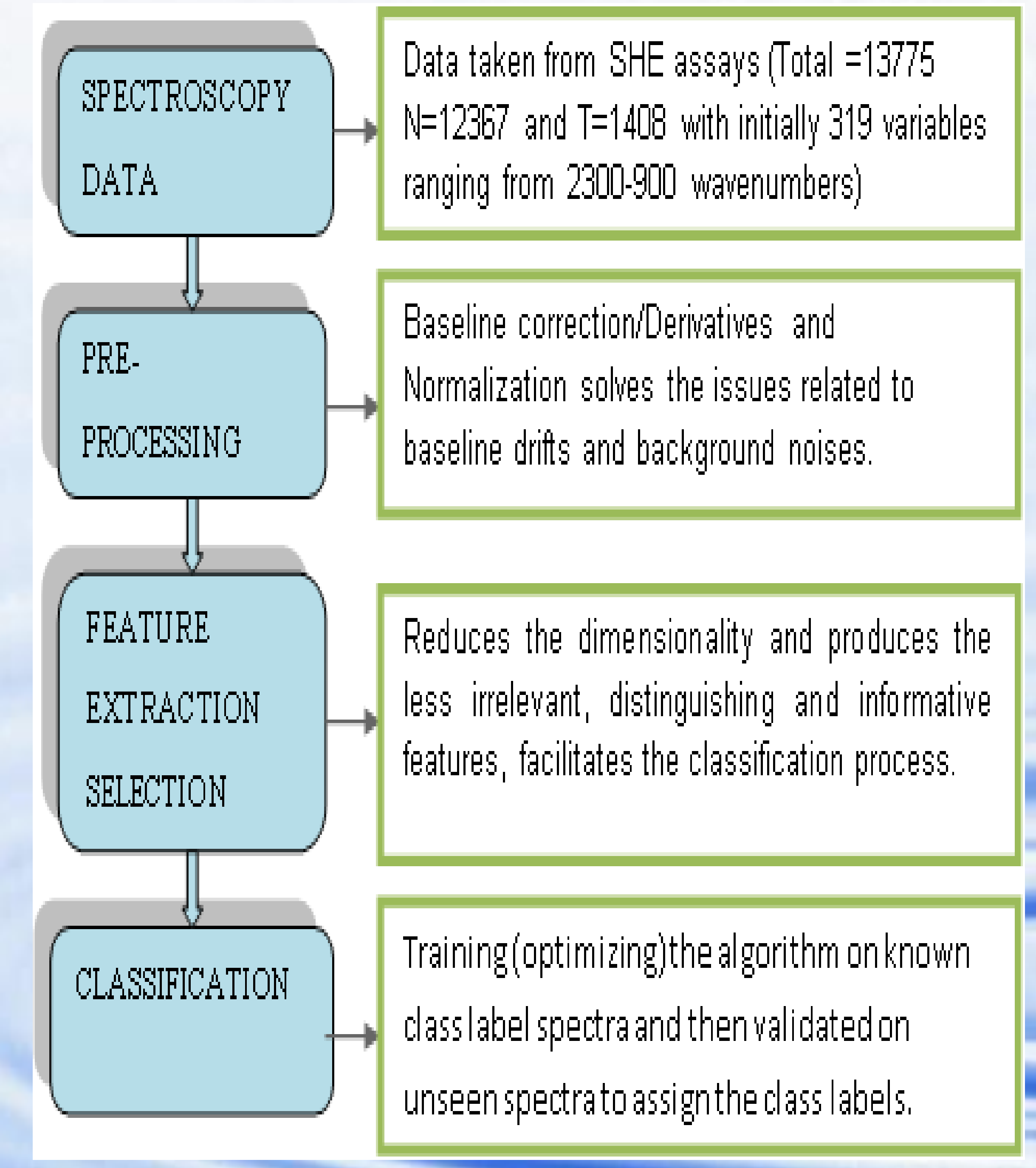
- ✓ Interpretation is objective
- ✓ Minimally skilled technician

The Biological molecules, when exposed to radiation in the mid IR region, Characteristic absorptions from the excitation and vibration of bonds with in the molecules can be exhibited in the form of spectrum.



This work is carried out under collaboration of biology department and Department of Communication systems with the company Uniliver, USA.

Classification Strategy



SPECTROSCOPY DATA → Data taken from SHE assays (Total =13775 N=12367 and T=1408 with initially 319 variables ranging from 2300-900 wavenumbers)

PRE-PROCESSING → Baseline correction/Derivatives and Normalization solves the issues related to baseline drifts and background noises.

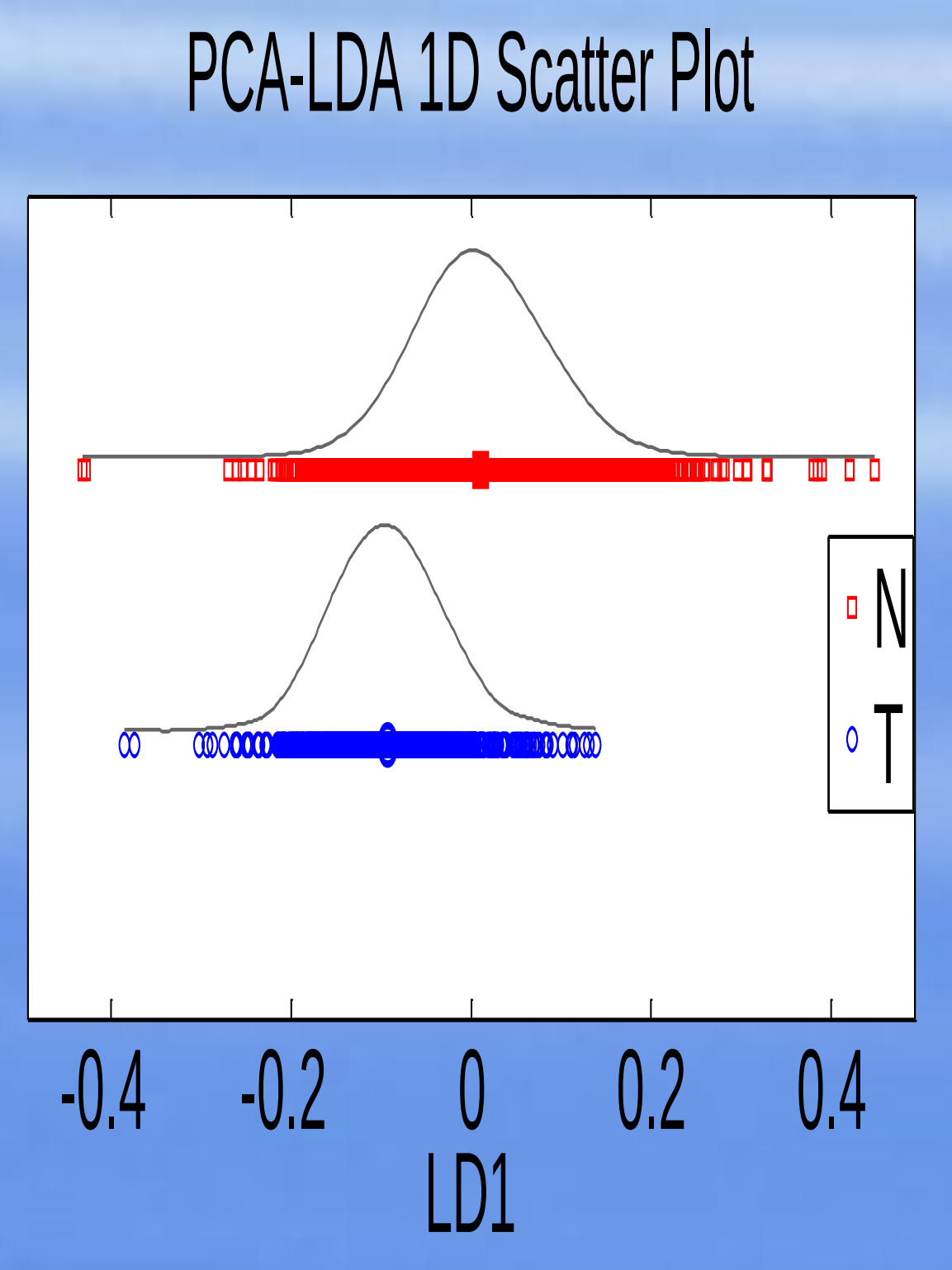
FEATURE EXTRACTION SELECTION → Reduces the dimensionality and produces the less irrelevant, distinguishing and informative features, facilitates the classification process.

CLASSIFICATION → Training(optimizing)the algorithm on known class label spectra and then validated on unseen spectra to assign the class labels.

PCA-LDA

The goal is to find a small set of uncorrelated variables which accounts for most of variance of the original data set and are suitable for classification. Each new variable is a linear combination of original variables (wavenumbers)

LDA is a supervised technique whereby the outputs from PCA are linearly-combined in order to maximize the ratio of between-class and within-class variances



Classifiers

The task of classification is to predict unknown data based on mathematical models derived from training data. Classifiers should be reliable in the sense of having high generalization power

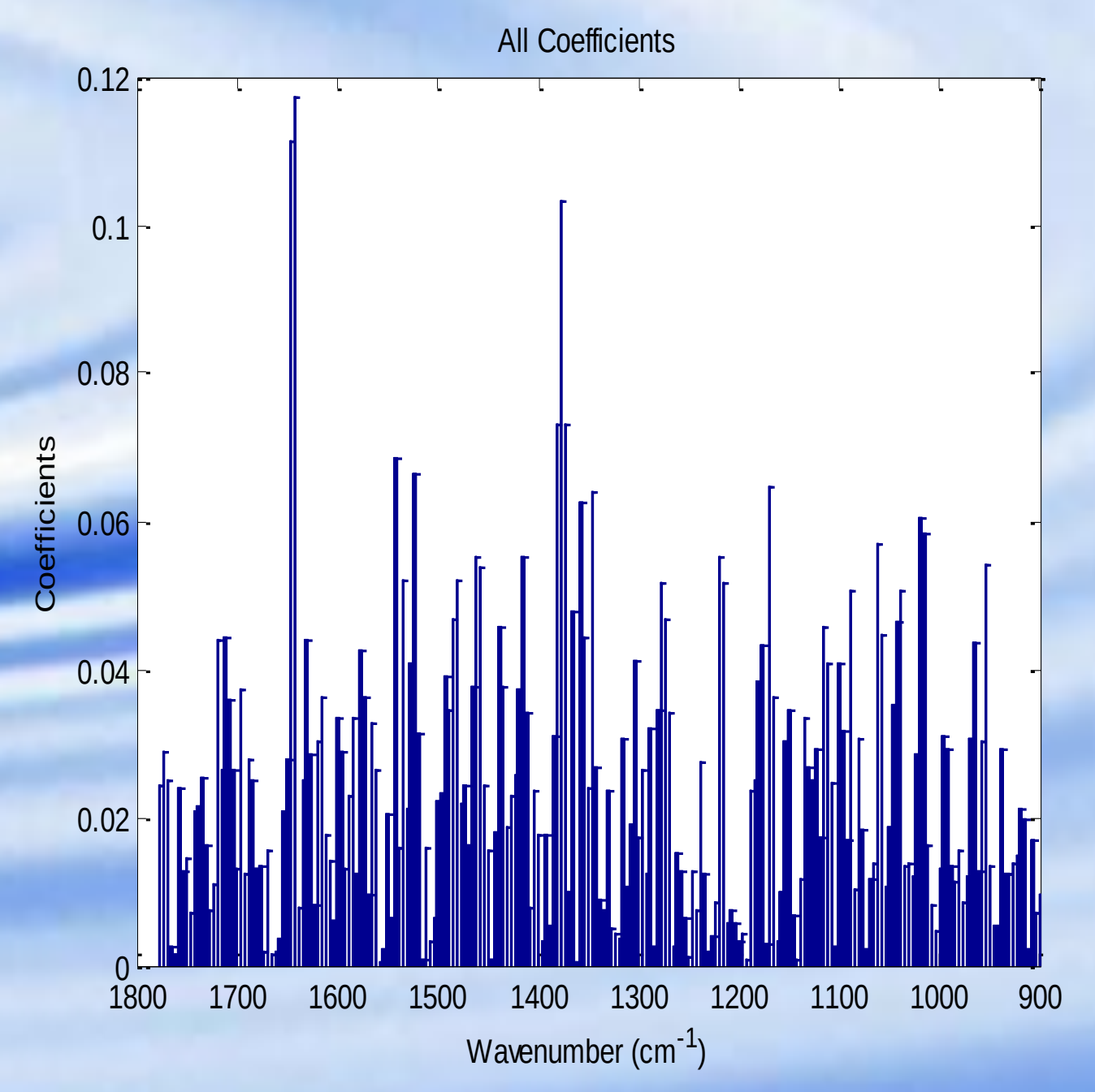
Evolving Fuzzy Models

eClass is a self developing (evolving) fuzzy rule based classifier system based on the Takagi-Sugeno type. In the training stage the structure of the model change (evolve) and its parameters adapt gradually to every new sample. The linguistic structure of the rule is shown as:

$$R^i: IF\left(x_1 \ is \ around \ x_1^{i*}\right) AND \ldots AND\left(x_n \ is \ around \ x_n^{i*}\right)$$
$$THEN \left(q_1^i = x^{-T}\alpha_1^i\right) AND \ldots . AND\left(q_C^i = x^{-T}\alpha_C^i\right)$$

An important characteristics of eClass is the ability to select the features (wavenumbers) from the large set of initial variables based on the accumulated contributions.

The value of these weights corresponding to each weave-number is very important regarding the analysis of biochemical composition of the cells. Weights are used for the gradual removal of the wave-numbers (25 selected)



Distance To Weighted Mean(DTWM)

The classifier uses the density function to weight the data points and compute the minimum distance from the weighted mean instead of traditional mean. The recursive density estimation, its internal coefficients and thus the weighted mean itself can be updated for every new sample. Weighted mean can be more robust against noisy conditions and non-uniform spread of data

Potential: $P_c(v) = \dfrac{N_c}{(p^T p+1)-2p^T\xi_{ck}+\beta_{ck}}$

Weighted Mean:

$$\overline{x_c} = \frac{1}{\sum_{i=1}^{N_c} P_c(x_i)} \sum_{i=1}^{N_c} P_c(x_i) \times x_i$$

Results

Some well known algorithms such as K-nn is also tested along with MATLAB build in Classifiers: Quadratic Discriminant Analysis(QDA) .

Sr #	Data type	Classifiers				
		DTWM	eClass	QDA	Knn	1-Rule-per-Class
1	BC Data 229 features	81.53 *[PCA+ LDA]*	*65.25* *[PCA]*	75.05 *PCA*	*62.53* *[LDA]*	81.58 *[PCA+ LDA]*
2	BC Data 25 features	82.68 *[PCA+ LDA]*	65.52	71.44 *[PCA]*	59.83 *[LDA]*	81.74 *[PCA+ LDA]*

Conclusions

- The results that have been achieved from these comparisons are highly encouraging as the weighted mean approach and 1-Rule per class fuzzy model outperformed in classification accuracy for both type of data sets.
- PCA-LDA has shown a huge potential as a feature transformation technique.
- Feature selection using eClass is important not only because it has improved the model performances, but also it can help in understanding the underlying mechanism of the disease.

Email Author: s.khan1@lancaster.ac.uk