Correlation Based Feature Selection for Proteomics

Qian Zhang
School of Computer Science, University of Birmingham
msc35qxz@cs.bham.ac.uk

Introduction:

As a kind of preprocessing technique commonly used on high-dimensional data, feature selection reduces the dimension, removes the irrelevant and redundant features, reduces the amount of data needed for learning, improves algorithm's predictive accuracy and increases the constructed model's comprehensibility without altering the original representation of the variables. During the last decade, the feature selection techniques have became a real prerequisite for model building in bioinformatics.

This paper focuses on the mass spectra analysis to emerge a new and attractive framework for disease diagnosis and proteomic profiling. The data analysis step is severely constrained by both high dimensional input spaces and their inherent sparseness, just as it is the case with gene expression datasets.

By using the correlation based feature selection method, the experiment aims to :

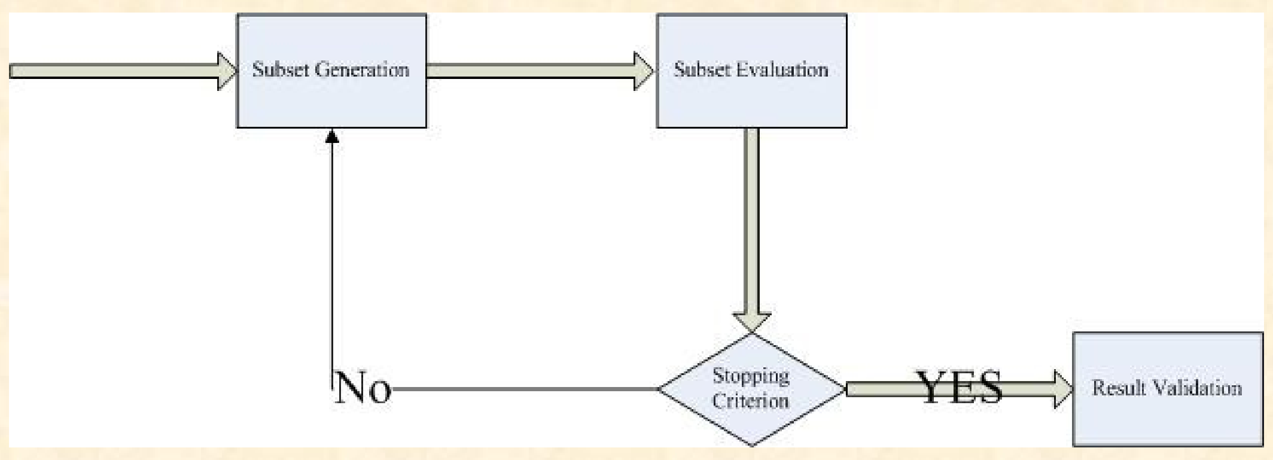
Starting from the raw data, and after an initial step to reduce noise and normalize the spectra from different samples, the extracted variables that will constitute the initial pool of candidate discriminative features are required.

Perform aggressive feature extraction procedures using elaborated peak detection and alignment techniques.

A correlation based filter approach and the embedded capacity of random forests algorithm constitute the feature selection strategy.

Feature Selection:

A typical feature selection process consists of four basic steps as follow, our approach is extended on these four steps:



Correlation Based Measures:

A feature is good if it is relevant to the class concept but is not redundant to any of the other relevant features. Applied with correlation, the goodness of feature is measured whether it is highly correlated with the class but not highly correlated with any of the other features. The information theoretical concept of entropy is introduced as the measurement of the uncertainty of a random variable. It is defined as :

$H(X) = -\sum P(xi)\log2(P(xi))$; $H(X|Y)= -\sum P(yi) \sum P(xi|yj)\log2(P(xi|yj))$;

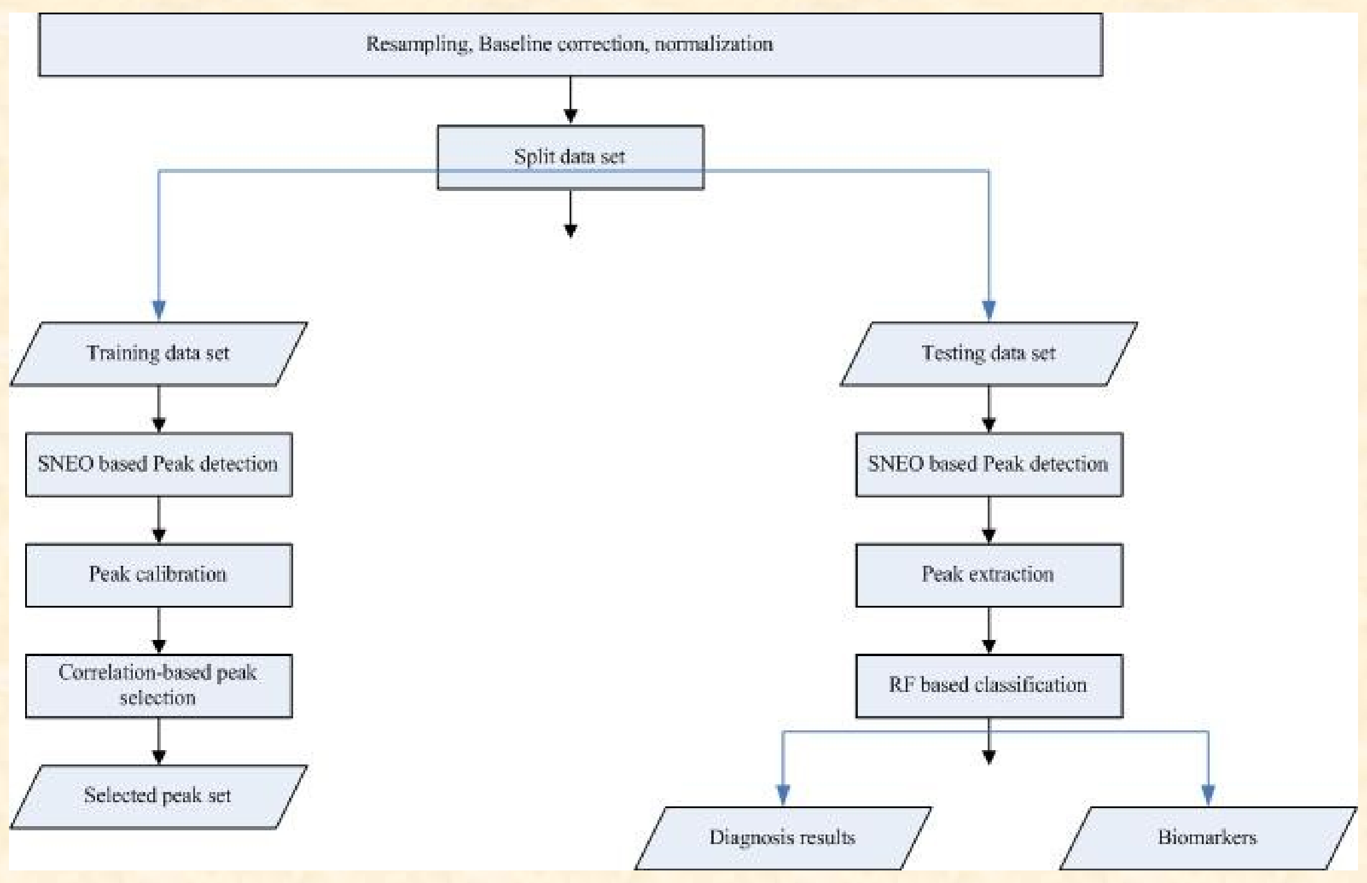
$IG(X|Y)=H(X) -H(X|Y)$;

$SU(X,Y)=2[IG(X|Y)/(H(X)+H(Y))]$

Process Description:

Main steps: 1) Data preprocessing: normalization the data and split the data set into training and testing two sets ; 2) SNEO based peak detection in both sets; 3) Select the training peak set using correlation based method based on the calibrated peak; 4) Apply the result get from last step on testing date to extract features as input for next step; 5) Random forest classifier generates the results

The details are shown in following figure:



Results:

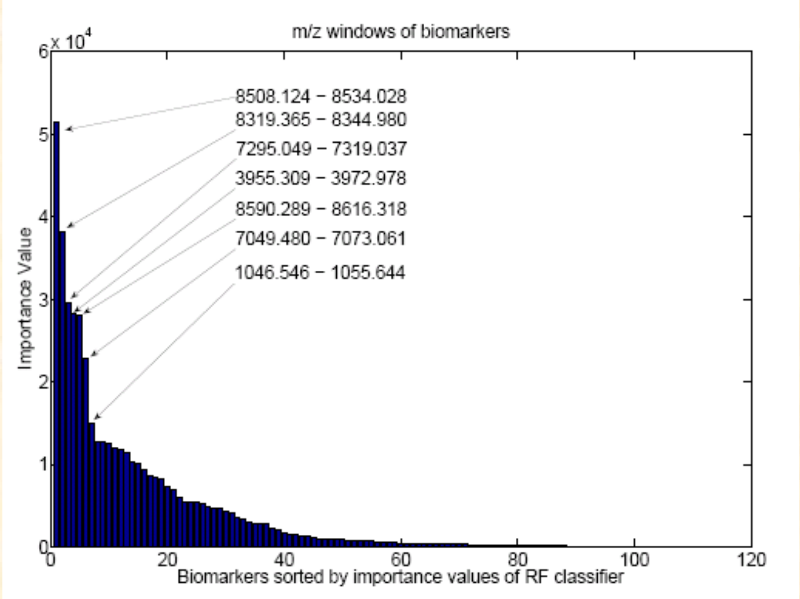
The results gotten are compared with other two proposed methods:

3.Yasui: the peak is detected if it takes the maximum value in the k-nearest neighborhood.

4.Cromwell package: use a Discrete Wavelet Transform (UDWT) first and then peak detected by locating maxima in each proposed spectrum and then are consequently qualified with Signal-to-Noise ratios. Finally, the detected peaks are calibrated by combining peaks that differed in location by no more than 7 clock ticks.

Repeated the whole procedure of the three methods for 1000 times and calculated the average results, it can be found that the overall test set accuracy generated by the correlated based selection are all better than the Yasui and Cromwell's methods.

After feature selection, biomakers sorted by feature importance values generated by random forest classifer shown in fig:



Discussion:

The core of this approach is the correlation based filter selection which achieves the high level of dimensionality reduction by selecting the least number of features and for most of the data sets, it maintain or even increase the accuracy.

The peak selection is employed to select a parsimonious peak set that generates the most accurate classification results and RF classifier is then applied to identify the prediction on the selected peak set.

It obtains more biologically meaningful results for further study and validation.

Reference:

Y. Saeys, I. Inza and P. Larranaga(2007). A review of feature selection techniques in bioinformatics. (Augutst 24,2007). Bioinformatics Advance Access published.

S. He and X.L. Li. Profiling of high-throughput mass spectrometry data for ovarian cancer detection. Cercia, School of Computer Science, The University of Birmingham.

L. Yu and H. Liu(2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. (2003), Proceeding of the Twentieth International Conference on Machine Learning(ICML-2003).

Acknowledgements:

I would like to express my gratitude to all those who gave me the possibility to complete this dissertation. I am deeply indebted to my supervisor Dr. Shan He whose help, stimulating suggestions and encouragement supported me throughout the first mini-project and the final project.