Feature Selection & Hybrid Decision Tree/Genetic Algorithm for Cancer Diagnosis based on Mass Spectrometry Proteomics

Henri F. Tedom Noumbi
School of Computer Science, University of Birmingham
msc45hft@cs.bham.ac.uk

Introduction

Applying data mining techniques enable to identify trends within the data that people did not know existed, and leverage the data to create new opportunities or values for organizations, as they provide timely and accurate information for decision-making purposes.
Medicine and bioinformatics in particular are benefiting greatly from the advances in the discipline. As an example, 23.1% of all death in the US in 2004 was caused by cancer (US centre for statistics on cancer), whilst ovarian cancer accounts for 4% of all female cancers, 90% of those women could be saved if the disease was detected at stage I [1].
Mass spectrometry as a proteomic tool has recently been applied to early-stage cancer diagnosis. It enables the creation of individual's profiles, which analysed together can reveal the underlying patterns that govern the formation of cancers. Unfortunately, these profiles are in very high dimensions, and because of the "curse of the dimensionality" can be tricky to manipulate.

This work proposes a novel profiling method that combines the Smoothed Nonlinear Energy Operator (SNEO), the Nearest Shrunken Centroid (PAM) for features selection, the Decision Tree (DT) classification, and the Genetic Algorithm (GA), to find a parsimonious set of biologically meaningful biomarkers and rules that explain the formation of anomalies.

The Method

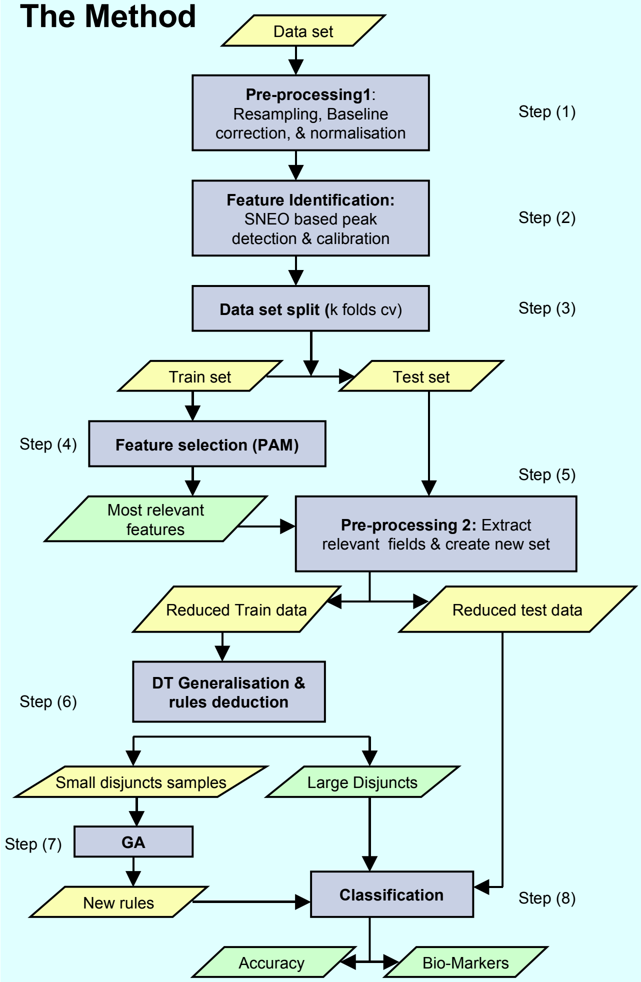
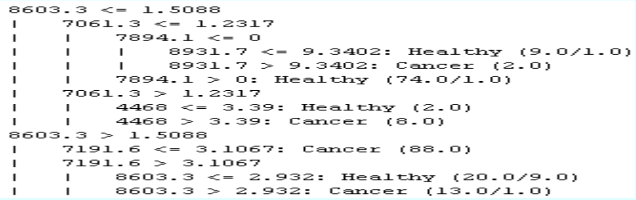


Figure 1. the different stages of the proposed technique

We used mainly the ovarian cancer data (high & low-resolution). The first two steps correspond to the 1st three of the method proposed in [1]. PAM uses the idea of LDA and within-class standard deviation to compute the standardized centroid for each class (cancer, or healthy). Each class centroid is then shrunk by an amount called threshold (decided by c-v) towards the overall centroid for all classes. All the non-zero peaks remaining are the most relevant features.



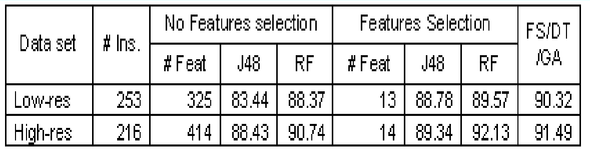
```
8603.3 <= 1.5088
|    7061.3 <= 1.2317
|    |    7894.1 <= 0
|    |    |    8931.7 <= 9.3402: Healthy (9.0/1.0)
|    |    |    8931.7 > 9.3402: Cancer (2.0)
|    |    7894.1 > 0: Healthy (74.0/1.0)
|    7061.3 > 1.2317
|    |    4468 <= 3.39: Healthy (2.0)
|    |    4468 > 3.39: Cancer (8.0)
8603.3 > 1.5088
|    7191.6 <= 3.1067: Cancer (88.0)
|    7191.6 > 3.1067
|    |    8603.3 <= 2.932: Healthy (20.0/9.0)
|    |    8603.3 > 2.932: Cancer (13.0/1.0)
```

Figure 2: An example of classification tree

The Characteristics of the GA:
- ● Two different variations: GA-Small and GA-Large-SN
- ● Representation: Fixed length chromosome with No. Of attributes number of gene as antecedent of the rule
- ● The fitness function is sensibility x specificity; standard one point-crossover 80%; 1% mutation; selection tournament with size depending on variation, as the population size, & the No. of generation; elitism factor 1; rule pruning operators; & threshold according to size of the data set.

Results

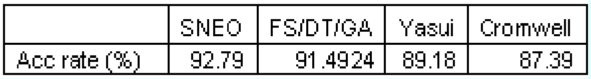
To compare our results with that of others on this benchmark problem, we used 2-folds cross validation, repeat the experiment 5 times, and compute the average of the outcomes. Figure 3 shows the results obtained:



Data set	# Ins.	No Features selection			Features Selection			FS/DT /GA
		# Feat	J48	RF	# Feat	J48	RF	
Low-res	253	325	83.44	88.37	13	88.78	89.57	90.32
High-res	216	414	88.43	90.74	14	89.34	92.13	91.49

Figure 3

Comparison with the results obtained in [1] on the high resolution data:



	SNEO	FS/DT/GA	Yasui	Cromwell
Acc rate (%)	92.79	91.4924	89.18	87.39

Figure 4

Discussion

The hybrid FS/DT/GA uses a number of established algorithms. It promotes their individual strengths, whilst constraining and cancelling each other's short comings. PAM deals well with the high dimensionality of the data, but cannot provide the rules that explain the effect of the biomarkers involved. The DT provides such knowledge but suffers from the effects of small disjuncts, which can be well dealt with by the GA, whom search space is constrained by the FS and DT.

There are algorithms known to perform better in term of accuracy than our method, but they sacrifice the biological interpretability [1], which is equally important as it enable to target and plan a specific course of action (here the treatment). However we offer the best trade off between accuracy and meaningfulness.

Furthermore, and unlike SNEO where a different set of relevant features can be found using the same set of samples for different runs, we found no such inconsistencies and even fewer numbers of relevant features.

Acknowledgements & References

My gratitude is extended towards my supervisor Dr. Shan He, not only for his guidance and support, but also for providing the pre-processed data.

[1] Shan He, and Xiaoli Li; Profiling of high-throughput mass spectrometry data for ovarian cancer detection; Cercia, School of Computer Science, The university of Birmingham; Birmingham; UK; 2007.