Genetic Programming Applied to the Rapid Spectroscopic Analysis of Biological Samples

Janet Taylor
Dept. Computer Science
University of Wales,
Aberystwyth
Ceredigion SY23 3DB,
United Kingdom
jjt95@aber.ac.uk

Jem J Rowland
Dept. Computer Science
University of Wales,
Aberystwyth
Ceredigion SY23 3DB,
United Kingdom
jjr@aber.ac.uk

Douglas B Kell
Biological Sciences
University of Wales,
Aberystwyth
Ceredigion SY23 3DD
United Kingdom
dbk@aber.ac.uk

ABSTRACT

Spectroscopic techniques enable rapid, highly characteristic fingerprinting of biological samples in classification and quantification applications. However the size and complexity of the data sets do not easily allow the formation of a concise, explicit model using standard data analysis methods. Initial studies show that 'classical' GP can provide solutions comparable with standard methods in terms of predictive ability, but the models formed can be complex. A hybrid evolutionary system is described which constrains the output expression yet provides an accurate, interpretable model to relate spectral features to (bio)chemical features of samples under investigation.

1. Initial Studies

Initial investigations into the applicability of genetic programming (GP) (Koza, 1992) included classification of bacterial strains on the basis of the normalised output of a spectral analysis technique (Pyrolysis Mass Spectrometry) (Taylor *et. al.*, 1998a) This, and other experiments (Gilbert *et. al.*, 1997; Jones *et. al.*, 1998; Taylor *et. al.*, 1998a) using the GP method have proved comparable with current methods of biological data analysis in terms of predictive accuracy. Studies using Fourier transform infrared (FT-IR) spectroscopy have also shown the tolerance of GP to noise in biological data, leading to a reduction in the need for pre-processing (Taylor *et. al.*, 1998b). The GP solutions enable input variable selection, but are generally complex. This led to investigation of the use of constrained output expressions.

2. Current Research

In the developing system the form of the output expression is highly constrained so as to enable clear input variable selection rules to be determined. Many of the datasets under examination are quasi-continuous in nature with 882 or more input variables. The high number of candidate variables and the quasi-continuous nature of the data suggested that the use of a constrained form of mutation would provide an accelerated search of the selection space. A population of continuous regions of variables, encoded by their position in the data and the region size, is mutated to 'scan' the dataset to identify a combination of significant variables that identify a relationship in the data. To constrain the output expression, these regions are weighted and combined in an expression, similar in structure to the chromosome in a genetic algorithm (Goldberg, 1989) with a fixed length genome but with integer representation. Experiments with similar data have shown that basic arithmetic operators in the function set are sufficient to form the equation, but with genetically selected powers so as to provide further expressive ability. The fitness function decodes these individual regions of variables into their averages, evaluates the resulting expressions, and scores the individuals according to the root mean squared error (RMSE). This form of expression facilitates interpretation, not only to identify significant variables, but to understand their relationship to the original spectroscopic data.

References

Koza, J. R., *Genetic Programming: On the Programming of computers by Means of Natural Selection* (MIT Press, Cambridge, MA, 1992).

Taylor, J., *et. al.*, *FEMS Microbiology Letters* **160**, 237 - 246 (1998a).

Gilbert, R. J., *et. al.*, *Anal. Chem.* **69**, 4381 - 4389 (1997).

Jones, A., *et. al.*, *Biotechnology and Bioengineering in Press* (1998).

Taylor, J., *et. al.*, Genetic Programming in the Interpretation of Fourier Transform Infrared Spectra: Quantification of Metabolites of Pharmaceutical Importance, J. R. Koza, et al., Eds., Genetic Programming 1998, Madison, Wisconsin, USA (Morgan Kaufmann, 1998b).

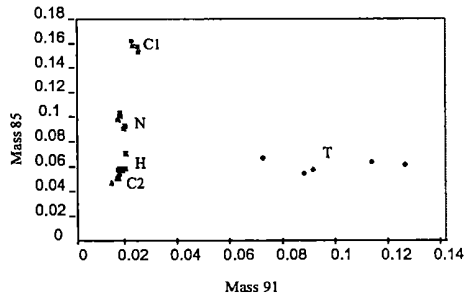
Goldberg, D. E., *Genetic Algorithms in search, optimization and machine learning* (Addison-Wesley, 1989).

Category: Genetic Programming
Ph.D. Thesis Title: Genetic Programming Applied to the Rapid Spectroscopic Analysis of Biological Samples
Author: *Janet Taylor
Supervisors: *Mr Jem J Rowland and †Professor Douglas B Kell
*Dept. of Computer Science and †Institute of Biological Sciences, University of Wales, Aberystwyth, SY23 3DA UK
Email jjt95/jjr/dbk@aber.ac.uk Tel: +44 (0)1970 623111

Biological data can be modelled by various statistical methods (1) which are prone to failure where relationships between variables are complex and highly non linear in nature. Artificial Neural Networks (2) may be used as a supervised non linear modelling method which, whilst capable of giving accurate predictions of quantity or quality, do not readily divulge the process of the data manipulation involved in the deduction of a suitable model.

This work is investigating Genetic Programming (3) as a supervised learning tool for the modelling of complex data such as these. The model produced by a GP is visible as the expression produced. Significant variables in the training data set may be identified from these output expressions and thus the method is able to give insight into the chemical features that distinguish between the training samples.

An example of the initial investigations into the applicability of GP was a classification, on the basis of the normalised output of a spectral analysis technique (Pyrolysis Mass Spectrometry), of four different known strains of a bacterial species (designated as T, C1, C2 and N) and one unknown hospital isolate strain (designated H) (4). The training of 4 small populations, each consisting of 15 individuals resulted in expressions selecting different variables for each strain. Upon examination of the output expression three variables from the original input data were identified as being particularly important, each of which may be related to different chemical features. Plotting just two of these variables results in tight clustering (see figure). This, and other experiments (5) using the classical Genetic Programming method have proved comparable with the current methods of biological data analysis in terms of predictive accuracy.



We have also shown the tolerance of Genetic Programming to varying sources of noise in biological data, leading to a reduction in the need for pre-processing. Our current aim is to modify the classical Genetic Programming method to constrain the structure and the complexity of the output expressions to provide an easily interpretable expression in terms of combinations of variables (6, 7). This will aid the interpretation of the output for variable selection and relationship to biological (or biochemical) features. Many of the datasets under examination are quasi-continuous in nature with high dimensionality (greater than 882 input variables). The high number of candidate significant variables suggests that an emphasis shift from crossover to a constrained form of mutation would provide an accelerated search of the selection space. A population of continuous regions of variables, encoded by their position in the data and the size of the region will be mutated to 'scan' the dataset to identify a combination of significant variables that identify a relationship in the data. To constrain the output expression, these regions will be weighted and combined in an expression, similar in structure to the chromosome in a Genetic Algorithm, but with a variable length genome and with integer representation. Experiments with similar data have identified the need only for basic arithmetic operators in the function set to form the equation, but with genetically selected powers so as to provide further expressive ability. The fitness function will necessarily be a complex operation, to decode the region of variables into the average or median of the values contained within, fit the resultant value into each equation and then evaluate the expression, and score the individual. This will be compensated for by the ease of interpretation of each candidate expression, not only to identify the significant variables, but to clearly understand the relationship between them.

The emphasis of this hybrid system is to evolve accurate prediction models of quasi-continuous data, initially focusing on Fourier Transform -Infrared (FT-IR) spectra and the quantitative determination of constituents within biological samples, through the formation of models which will be easily interpreted in terms of the original spectrum and the underlying chemical structure.

1. H. Martens, T. Næs, *Multivariate calibration* (John Wiley, Chichester, 1989).
2. R. Goodacre, et al., *FEMS Microbiology Letters* **140**, 233-239 (1996).
3. J. R. Koza, *Genetic Programming: On the Programming of computers by Means of Natural Selection* (MIT Press, Cambridge, MA, 1992).
4. J. Taylor, R. Goodacre, W. Wade, J. Rowland, D. Kell, *FEMS Microbiology Letters* **Submitted** (1997).
5. R. J. Gilbert, R. Goodacre, A. M. Woodward, D. B. Kell, *Analytical Chemistry* **69**, 4381 - 4389 (1997).
6. R. P. Paradkar, R. R. Williams, *Applied Spectroscopy* **51**, 92-100 (1997).
7. Z. Michalewicz, *Statistics and Computing* **4**, 141 - 155 (1994).