Using an Adaptive Fuzzy Logic System to Optimise Knowledge Discovery in Proteomics

James Malone, Ken McGarry and Chris Bowerman
School of Computing and Technology
Sunderland University
St. Peter's Way, Sunderland, SR6 0DD, United Kingdom
e-mail: {james.malone, ken.mcgarry, chris.bowerman}@sunderland.ac.uk

Abstract*: The growth of biomedical databases has seen a demand for data mining techniques to efficiently and effectively analyse the data contained within. One important consideration is the need to include expert's opinions within the knowledge discovery process. However, this can be difficult to accomplish when such heuristics are presented in loosely defined, fuzzy terms. We present the use of an Adaptive Nero-Fuzzy Inference System (ANFIS) as an approach to optimizing such fuzzy opinions with the long term aim of incorporating such optimized rules into goal and data driven data mining of proteomic's data*.

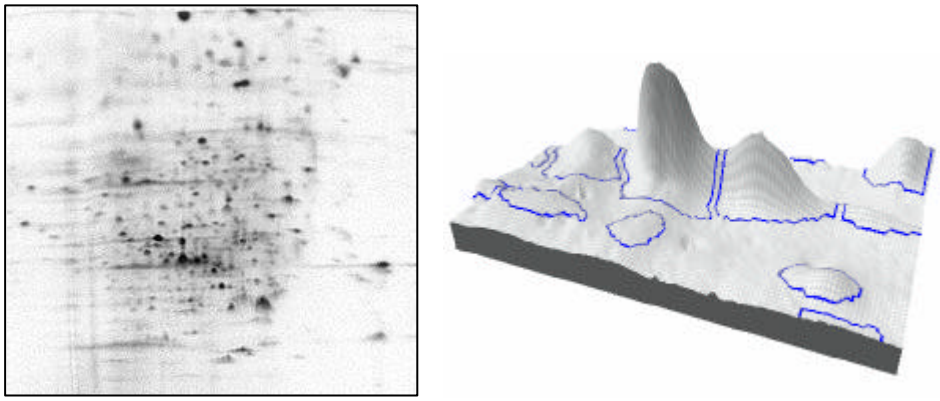
Keywords*:* ANFIS, fuzzy opinions, data mining, proteomics data, 2-DE gel analysis

1. Introduction

In the last decade, proteomics has grown rapidly as an area of research, largely driven by technological improvements, and is often described as the next step to dramatically advance drug discovery [1]. Within proteomics, two-dimensional electrophoresis (2-DE) is unrivalled as a technique to analyse protein expression [2, 3] and is a key component of current proteomics research [4]. 2-DE is capable of separating thousands of proteins according to their electrical charge and molecular weight. Such experiments can create large amounts of high-dimensional, spatio-temporal experimental data. Using staining techniques, these proteins appear as spots on these gels (as shown in Figure 1). The gel images are digitised by image processing and analysis software, such as the popular "Progenesis" product which produced the images in Figure 1.

Analysis of the data generated from such 2-DE gel experiments has been identified as the most important task in proteomics [5]. However, discovering information from this data is a complex and difficult task because of the large number of spots, variations in shape location and size [6]. This problem is compounded when considering that experiments are often conducted over a series of gels, hence, the aim is detecting trends across a large number of images. At present, the analysis of 2-DE gel data is normally conducted manually which is both time consuming and requires considerable expertise.

Further development of database analysis algorithms would be of huge benefit to proteomics [7] since current methods are unable to handle and analyse such results meaningfully [8]. Since current analysis is conducted by experts using heuristic knowledge gathered over years of experience within the area, an important consideration would be the incorporation of such expert domain knowledge within any automated analysis method devised. One issue relating to this expert knowledge is that expert's opinions vary greatly from institution to institution. Following knowledge acquisition conducted on 20 experts, results presented varying opinions on metrics which are used to identify 'interesting' protein characteristics. This presents problems when attempting to extract optimum values for such metrics for incorporation into goal driven data mining for trend analysis purposes.



(a) (b)

Figure 1. (a) 2-DE Gel Image, protein spots appearing as black dots on the image (b) a section of the gel showing several protein spots, describing the 3 dimensional aspect of the gel

We investigate the use of an adaptive fuzzy logic controller, specifically Adaptive Nero-Fuzzy Inference System (ANFIS) [9] to optimise the fuzzy if-then type rules obtained from human experts. The long term aim of such an investigation is to incorporate these optimised rules into goal driven data mining to identify interesting proteins from 2-DE gel data. The remainder of the paper is structured as follows. Section 2 describes the ANFIS model with a break down of each layer, section 3 describes the Proteomic's data used and input parameters to the model and presents results of the training as well as an example of some of the optimised fuzzy rules extracted following this process. Finally, conclusions are drawn in section 4.

2. An adaptive nero-fuzzy inference system

ANFIS is a fuzzy inference system implemented within the architecture and learning procedure of adaptive networks [9, 10, 11]. An adaptive network is a superset of all kinds of feed-forward neural networks with supervised learning capability. ANFIS can be used to optimise membership functions to generate stipulated input–output pairs and has the advantage of being able to subsequently construct fuzzy "if–then" type rules representing these optimised membership functions. There are various successful examples of ANFIS used in biomedical applications [12, 13, 14].

2.1 ANFIS Architecture

The ANFIS structure used in this paper consisted of a 5-layer Sugeno type architecture, a typical example of which is seen in Figure 2. In this example, two inputs are used (*x*,*y*) and one output (*f*) (which is a limitation of Sugeno-type systems, i.e. that there is only a single output, obtained using weighted average defuzzification (linear or constant output membership functions)).

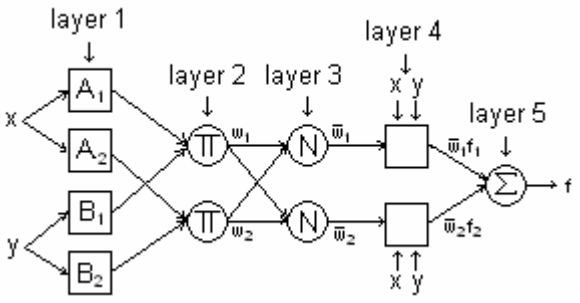


Figure 2. A 5-layer ANFIS structure [9]

In the first layer, all nodes are adaptive, i is the degree of the membership of the input to the fuzzy membership function (MF) represented by node;

$$O_{1i} = \mu_{Ai}(x) \qquad i = 1, 2 \qquad (1)$$

$$O_{1i} = \mu_{Bi-2}(y) \qquad i = 3, 4 \qquad (2)$$

where O_{li} is the output of the node i in a layer l.

In the second layer the nodes are fixed (i.e. not adaptive). Nodes in this layer are labelled Π and multiply the signal before outputting. The outputs are given by;

$$O_{2i} = w_i = \mu_{Ai}(x)\ \mu_{Bi}(y) \qquad i = 1, 2 \qquad (3)$$

Each node output in this layer represents the firing strength of the rule.

In the third layer, every node is also fixed and are labelled with an N and perform a normalisation of the firing strength from the previous layer. The output of each node is given by;

$$O_{3i} = \bar{w}_i = \frac{w_i}{w_1 + w_2} \qquad i = 1, 2 \qquad (4)$$

In the fourth layer, all nodes are adaptive. The output of a node is the product of the normalised firing strength and a first order polynomial and is given by;

$$O_{4i} = \bar{w}_i f_i = \bar{w}_i (p_i x + q_i y + r_i) \qquad i = 1, 2 \qquad (5)$$

where $\{p_i x, q_i y, r_i\}$ is the modifiable parameter set, referred to as *consequent parameters* since they deal with the *then* part of the fuzzy rule.

Finally, layer 5 is a single node labelled with Σ which indicates that the function is that of computing the overall output as the summation of all incoming signals;

$$O_{5i} = f = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \qquad i = 1, 2 \qquad (6)$$

Further details of the ANFIS model can be obtained from Jang [9] and Jang and Gulley [10].

3. Optimisation of Proteomics Heuristics

Following twenty knowledge acquisition sessions with experts from within the field of proteomics, heuristic knowledge was gleaned and represented in the form of fuzzy rules. Such rules have the advantage of being able to represent intuitive terms a form close to natural language and have no precise thresholds, reducing brittleness. These rules are to be used to provide a goal driven element to the final data mining process by allowing the incorporation of expert's opinions to identify what contributes 'interestingness' from within the proteomics data sets [15].

The structure used in this paper for experimentation consists of 6 inputs, described in Table 1. These inputs form the basis of the experts opinions and are used in various rules to describe whether or not a particular protein spot is 'interesting' or not and hence worthy of further laboratory analysis.

Table 1. Inputs used in ANFIS structure

Input	Description	No. Membership Functions
Absence/Presence	The absence and presence of protein spots from gel to gel	2
X/Y Movement	Movement of a protein spot in x and y dimensions from gel to gel	2
Volume Change	Increase or decrease in the volume of a protein spot from gel to gel	2
Budding	The joining or separation of protein spots from gel to gel	2
Shape Change	Morphological changes in shape, such as height from gel to gel	2
Percentage Variation	Percentage change in terms of spot abundance from gel to gel	3

3.1 Results

The data set was split into two; the training data set which was used to train the ANFIS model whilst a testing data set was used to verify accuracy of the trained ANFIS model. The training data set consisted of 75% of the total data set with the test data set consisting of the remaining 25%. Figure 3 shows the training of the ANFIS model, with the final convergence value approaching 0. The diagram shows that the error rate decreases rapidly over the first 25 epochs before quickly levelling out. Table 2 shows the classification accuracy of the ANFIS, with the accuracy for 'interesting' and 'non-interesting' proteins of 96% and 94% respectively.

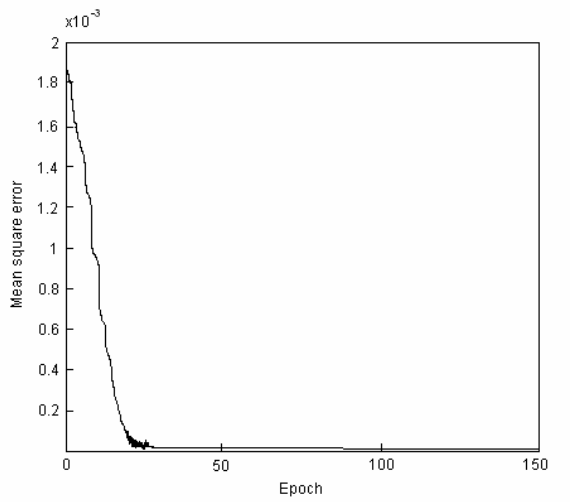


Figure 3. Network training error graph

Table 2. Test Results on ANFIS

Class	Classification Accuracy	Total No. Optimised Fuzzy Rules
Interesting Proteins	96%	14
Non-interesting Proteins	94%	82

Following the training and testing, the fuzzy expert rules were extracted and a sample of which is shown in Table 3. The membership functions of each input are described in fuzzy terms, such as high and low for all but %_Change which also has a medium membership.

Table 3. Optimised fuzzy rules

If (Absence/Presence is high) and (X/Y_Movement is low) and (Volume_Change is high) and (Budding is low) and (Shape_Change is high) and (%_Change is low) then (output is high)
If (Absence/Presence is high) and (X/Y_Movement is high) and (Volume_Change is high) and (Budding is high) and (Shape_Change is low) and (%_Change is low) then (output is high)
If (Absence/Presence is high) and (X/Y_Movement is high) and (Volume_Change is high) and (Budding is high) and (Shape_Change is high) and (%_Change is low) then (output is high)
If (Absence/Presence is high) and (X/Y_Movement is high) and (Volume_Change is high) and (Budding is low) and (Shape_Change is low) and (%_Change is low) then (output is high)
If (Absence/Presence is high) and (X/Y_Movement is low) and (Volume_Change is high) and (Budding is low) and (Shape_Change is low) and (%_Change is high) then (output is high)

3.2 Knowledge Discovery from Fuzzy Rules

This optimisation process revealed which attributes appear to have most influence according to expert's opinions when classifying proteins as interesting or not. An analysis of the rules reveals that the Absence/Presence input parameter features as 'high' in most of rules that result in output as high (i.e. high level of interestingness). This was also true of the Volume_Change input which also featured as 'high' in most of the optimised fuzzy rules. It can also be noted from Table 2 that there were almost six times as many rules produced for non interesting proteins as there were for interesting. This seems to confirm the belief that only particular combinations of the input parameters result in interesting proteins according to the expert's beliefs. Although many of the expert's opinions in their initial form seemed to have varying biases towards particular attributes and at varying levels, the optimisation process has revealed that there is an underlying trend and hence agreement as to which attributes lead to interesting proteins. This can be shown in the high classification accuracy of the ANFIS model using these optimised membership functions that were derived from the initial fuzzy expert's opinions.

4. Conclusions

In this paper we have demonstrated the use of ANFIS to optimise expert's opinions. The ANFIS model offers the advantage of enabling use of initially approximate data in an effective manner whilst, following training, allowing fuzzy rules to be extracted which represent the optimised fuzzy membership functions. Such information is of particular use when considering goal-driven data mining techniques. The inclusion of expert's opinions into the data analysis process of proteomics data is seen as critical to extracting useful and interesting knowledge. Optimising the membership functions enables the use of previously messy expert's opinions to drive the data mining and hence improve the accuracy and value of any extracted results. Further work will concentrate on incorporating such optimised expert's opinions with data driven techniques to automatically analyse post-experimental proteomics data.

5. Acknowledgements

The authors acknowledge the support of EPSRC, Nonlinear Dynamics Ltd and Mark Elshaw.

References

[1] P.A. Whittaker (2003). What is the relevance of bioinformatics to pharmacology? In *Trends in Pharmacological Sciences*, vol. 24(8), pages 434-439, 2003.

[2] R.E. Jenkins and S.R. Pennington (2001). Novel approaches to protein expression analysis. In *Proteomics: From protein sequence to function*, pages 207-224 BIOS Scientific Publishers, Oxford, 2001.

[3] S.R. Pennington, S.R. Wilkins, D.F. Hochstrasser and M.J. Dunn (1997). Proteome analysis: from protein characterisation to biological function. In *Trends in Cell Biology*, vol. 17(4), pages 168-173, 1997.

[4] S. Beranova-Giorgianni (2003). Proteome analysis by two-dimensional gel electrophoresis and mass spectrometry: strengths and limitations. In *TrAC Trends in Analytical Chemistry*, vol. 22(5), pages 273-281, 2003.

[5] D.C. Liebler (2002). *Introduction to Proteomics: Tools for the New Biology*, Human Press, 2002.

[6] A. Thompson and T. Brotherton (1998). Information extraction from 2D electrophoresis images. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 20(2), pages 1060-1063, 1998.

[7] M. Mann and O.N. Jensen (2003). Proteomic analysis of post-translational modifications. In *Nature Biotechnology*, vol. 21(3), pages 255-261, 2003.

[8] M. Vihinen (2001) Bioinformatics in proteomics. In *Biomolecular Engineering*, vol.18, pages 241-248, 2001.

[9] J.-S.R. Jang (1993). ANFIS: Adaptive-network-based fuzzy inference systems. In *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23(3), pages 665-685, 1993.

[10] J.-S.R. Jang and N. Gulley (1995). *The Fuzzy Logic Toolbox for use with MATLAB*, The Mathworks Inc, 1995.

[11] J.-S.R. Jang, C.T. Sun and E. Mizutani (1996). *Neuro-Fuzzy and Soft Computing:AComputationalApproach to Learning and Machine Intelligence*, 1996.

[12] S.Y. Belal, A.F.G. Taktak, A.J. Nevill, S.A. Spencer, D. Roden and S. Bevan (2002). Automatic detection of distorted plethysmogram pulses in neonates and paediatric patients using an adaptive-network-based fuzzy inference system. In *Artificial Intelligence in Medicine*, Vol. 24(2), pages 149–165, 2002.

[13] I. Güler and E.D. Übeyli (2004) Application of adaptive neuro-fuzzy inference system for detection of electrocardiographic changes in patients with partial epilepsy using feature extraction. In *Expert Systems with Applications*, Vol. 27(3), pages 323-330, 2004.

[14] H.F. Kwok, D.A. Linkens, M. Mahfouf and G.H. Mills (2003) Rule-base derivation for intensive care ventilator control using ANFIS. In *Artificial Intelligence in Medicine*, Vol. 29(3), pages 185-201, 2003.

[15] Malone, J., McGarry, K. and Bowerman, C. (2004). Performing trend analysis on spatio-temporal proteomics data using differential ratio data mining. In *Proceedings of the 6th EPSRC Conference on Postgraduate Research in Electronics, Photonics, Communications and Software (Prep 2004)*, pages 103-105, 2004.