# Radial Basis Function Neural-fuzzy Model for Microarray Signature Identification

Julio De Alejandro Montalvo[1], George Panoutsos[1], Mahdi Mahfouf[1] and James W. Catto[2]
[1]*The University of Sheffield, Dept. of Automatic Control and Systems Engineering, Sheffield, U.K.*
[2]*The University of Sheffield, Academic Urology Unit, Sheffield, U.K.*

Keywords: Feature-selection, Neural-fuzzy, Fuzzy Logic, Radial-Basis-Function (RBF), Microarray, Bladder Cancer.

Abstract: This paper introduces a Fuzzy entropy-based method for the problem of feature selection. For the first time Fuzzy-Entropy is used to directly link the relative input relevance of a Radial-Basis-Function Neural-Fuzzy modelling structure. This embedded feature selection method uses the model performance as a criterion for the feature selection. The resulting model maintains its simplicity and transparency in the form of a linguistic Fuzzy-Logic rule-base. The proposed methodology is validated using a real biomedical case-study, which concerns the signature selection for the identification of the stage of bladder cancer. The signature selection and predictive modelling results are compared to previous research work on the same dataset, and it is shown that the RBF-NF model outperforms the previous modelling attempts by achieving high predictive accuracy (>90%). The model is shown to maintain its good performance even when using just 10 genes in the gene based signature.

## 1 INTRODUCTION

One of the biggest challenges in cancer research is the accurate early classification of tumours. This classification can reflect the stage of a tumour and may be achieved via a number of information sources, including clinical and radiological data and potentially, biochemical or molecular tests. However, limitations in the accuracy of these data have led to the search for more robust biomarkers such as gene expression data. One method for high throughput, global profiling of gene expression is the microarray. In this paper we investigate the ability for genetic data stage bladder cancer reliably. A reliable predictor capable of accurate classification at an early stage of the cancer would avoid unnecessary treatment and also save costs.

In recent years the study of microarrays has become more popular. Microarrays are chips that contain thousands of probes. These probes mirror the RNA or DNA sequence for individual genome locations. The expression or content of that corresponding genome structure can be measured by the abundance of binding to that probe. Microarrays have been used in a number of different contexts in human cancer.

Currently one of the most promising roles is as disease biomarkers that may be used to predict tumour stage and subsequent outcomes. Microarray data analysis is challenging due to the large size of these datasets(100s to 10,000 of probe values) and their imbalance with samples size (most series have less than 100 samples). This presents a challenging Systems Engineering classification and identification problem (high dimensionality, low number of samples). If accurate predictive models (wrapper methods) are built from the available gene data along with results from cancer biopsies and other clinical tests, one can then try to understand how the various genes and tests relate to cancer, and try to develop multi-dimensional patient prognostic maps that are capable of mapping cancer malignancy based on a minimum amount of data/tests.

The main challenges in this type of research are:

- The uncertainty of the data
- Model generalisation issues
- Identification of relevant genes/clinical markers
- Link model-based research findings with medical expertise

To address the problem of high number of input features, feature selection algorithms have become indispensable components of the learning process.

Feature selection is the process of detecting the


De Alejandro Montalvo J., Panoutsos G., Mahfouf M. and Catto J..
Radial Basis Function Neural-fuzzy Model for Microarray Signature Identification.
DOI: 10.5220/0004226801340139
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2013), pages 134-139
ISBN: 978-989-8565-35-8

relevant features and discarding the irrelevant ones. There are three categories for feature selection: filters, wrappers and embedded methods.

Statistical regression methods perform poorly when there are multiple interconnected variables and in the presence of contaminating data (Burke, Goodman et al., 1997). Filter and wrapper methods could be used in combination with Soft Computing (SC) techniques (i.e. Fuzzy Logic, Neural-Fuzzy Systems) to eliminate irrelevant genes at an early stage. The combination of SC with other computational techniques offers significant advantages in terms of imprecision tolerance and systems interpretability, and has proven to be an effective method performing equally or better than Support Vector Machines (SVM) or 'K Nearest Neighbourhood', which are very popular methods for gene expressions classification (Pal et al., 2007).

In this paper, a new embedded SC feature selection method is introduced based on Fuzzy Logic (FL) and a Radial-Basis-Function (RBF) Neural-Fuzzy (NF) computational structure. The presented methodology offers a feature selection that takes place during the model-training phase, whilst maintaining the system simplicity and interpretability. This is achieved by taking advantage of the Fuzzy Entropy measure (Al-Sharhan et al., 2001). Hybrid Neural-Fuzzy Logic models combine the learning ability of Neural Systems and the interpretability of Fuzzy Systems, they can automatically generate and adjust the membership functions and linguistic rules directly from the data. The presented method is a combination of Fuzzy C-Means and RBF-NF function; it is an embedded method as it trains the model while it performs the input selection. As a pre-input selection the t-test statistical method was used to reduce the large initial dataset. This is a popular pre-processing step in microarray gene selection, aiming at removing the irrelevant to the process genes. The proposed methodology, uses a variant of the the Levenberg-Marquardt algorithm for the model's parametric optimisation. The method is suitable for handling large datasets, and because of the 'IF-THEN' linguistic rules it helps the clinicians to understand how the model behaves.

The remainder of the paper is organised in four more sections as follows: 2. Radial Basis Function Neural-Fuzzy System: in this section a description of the modelling and data-mining structure is presented. 3. Fuzzy Entropy-Based Feature Selection: this is a detailed description of the new feature selection method 4. Gene Signature Selection: the new method is successfully applied to a bladder cancer literature dataset to predict the stage of the cancer, and finally, section 5: Conclusion and Future work.

# 2 RADIAL BASIS FUNCTION NEURAL-FUZZY SYSTEM

## 2.1 Clustering

The data-mining workflow consists of three stages, the first of which is Fuzzy C-means (FCM) clustering for the creation of the initial rule-base. This rule-base is then 'translated' into a Radial-Basis-Function Neural-Fuzzy structure, and is finally parametrically optimised via the Levenberg-Marquardt function-minimisation algorithm. The FCM method (Dunn, 1973) is frequently used in pattern recognition but the main reason to use it is because after Fuzzy C-Means is applied to a data it can be used directly as initial values of an RBF Model. FCM is based on minimisation of the following objective function:

$$J_{m=} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 , 1 \le m < \infty \qquad (1)$$

Where $m$ is any real number greater than 1, $u_{ij}$ is the degree of membership of $x_i$ in the cluster $j$, $x_i$ is the $i$th of d-dimensional measured data, $c_j$ is the d-dimension centre of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the centre.

## 2.2 RBF-Based Neural-fuzzy System

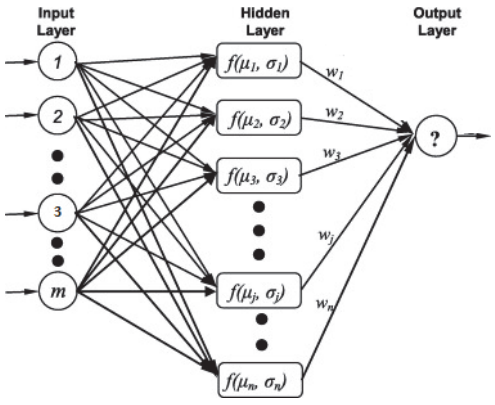


Figure 1: RBF Layers of NF model.

The second stage consists of applying the method proposed in (Panoutsos and Mahfouf, 2010). This

method uses an RBF function to describe a Neural-Fuzzy system. Figure 1 show the structure of the RBF-NF model, where the input, rule-base (hidden layer) and output layers can be identified. The presented system can then be parametrically optimised via a suitable function minimisation algorithm.

### 2.3 Levenberg-Marquardt Optimisation

The Levenberg-Marquardt (LM) algorithm is an iterative technique that locates the minimum of a multivariate function that is expressed as the sum of squares of non-linear real-valued functions (Levenberg, 1944). In this paper the RMSE between the training data and the model predicted data is used as the cost function to be minimised. The presented data-mining workflow provides an efficient and fast method for capturing numerical data-based information and converting it to a linguistic knowledge-base with a predictive capability.

The next section describes how the RBF-NF structure is exploited along with Fuzzy-|Entropy measures to identify relevant to the process features.

## 3 FUZZY ENTROPY-BASED FEATURE SELECTION

The presented method is based on two Fuzzy Logic features: The Fuzzy Entropy and the Tagaki-Sugeno-Kang (TSK) type (Takagi and Sugeno, 1985) of output layer for a NF system. Fuzzy-Entropy is a measure of 'fuzziness', it allows the quantification of how 'fuzzy' a value is when the Fuzzy Inference System (FIS) is used.The TSK output layer of an RBF-NF model is a linear combination of its inputs (polynomial).The hypothesis is that during model training, the values of the output weights $w_i$, for each rule, will increase (absolute value) for the inputs (genes) that are more 'influential' in (contribute to) the model predictions. One could analyse how the output weights change on every training iteration, hence determine the relevance of the corresponding inputs. This relationship in terms of entropy strength is relative to the genes, is measurable, and may be used to rank the genes for a particular rule in the rule-base. In the algorithmic process proposed here, the model is trained for 'N' iterations, while at 'n' iterations (n<N) the training can be 'paused' and the model can be reviewed in terms of the gene ranking order.

Not all the rules in the rule-base contribute with the same amount to the FIS. This is subject to the 'input space' of a particular gene. Therefore, the ranking order that may be established as a result of examining a single TSK rule is only relevant if the corresponding rule has a high contribution to the overall rule-base. This contribution can be established via the use of Fuzzy-Entropy (FE), as a measure of 'fuzziness'. The FE is calculated for each individual rule, and then a numerical 'index' is developed to 'adjust' the significance of the ranking of each individual rule. Finally, the overall ranking of the genes is calculated by using the FE-adjusted gene output weights.

In terms of the algorithmic process, Figure 2 summarises the gene feature selection. The first step is to rank the output weights by rule in descending order. The top 'n' genes are then selected and this information is passed on to the following step (this numerical threshold is process specific). The Fuzzy-Entropy is then calculated for each model prediction. The fuzzy entropy is defined using the concept of membership function. In 1972, De Luca and Termini defined Fuzzy Entropy Based on Shannon's functions and they introduced a set of properties for which Fuzzy Entropy should satisfy them (Al-Sharhan et al., 2001).

$$H_A = -K\sum_{i=1}^{n}\{\mu_i \log(\mu_i) + (1-\mu_i)\log(1-\mu_i)\} \quad (2)$$

Where:

$H_A = entropy, K = constant,$

$n = number\ of\ inputs,$

$\mu_i = membership\ degree$

Once the entropy is calculated, Eq. 3 is applied.

The 'B' index reflects the significance of a gene within a certain rule. The value of B is obtained for each gene in all the rules.

$$B = (\mu_i / H_A) * output\ Weight \quad (3)$$

The output weight is adjusted by the significance of a particular rule (proportional to the membership degree, inversely proportional to the 'fuzziness'). After the rule-adjusted significance per gene is calculated a new ranking order is then compiled.

The work presented in this paper is the first report, to our knowledge, of a Fuzzy-Entropy scheme applied to a RBF-NF modelling structure.

The resulting ranking of the genes directly relates to their performance in the modelling structure, is an iterative procedure – that can be

repeated a number of times as required – and provides a fast workflow to establish gene signatures from microarray data.

The dataset consists in 22,283 genes and 90 samples from the analysis of the Affymetrix Chip U133A Human Gene-chip. This study aims at predicting the 'Cancer Stage'.

# 4 GENE SIGNATURE SELECTION

The case-study presented in this paper is focused on the prediction of Bladder Cancer Stage using a dataset from a previous study made by Sanchez-Carbayo (Sanchez-Carbayo et al., 2006).

Table 1: Cancer stages.

| Value | Stage |
|---|---|
| 0 | PTA,PT1 |
| **1** | PT2, PT3A, PT3B, PT4, PT4A |

A common staging system uses numbers to indicate the stage of the cancer as shown in Table 1.

The cancer Stage values were 'encoded' to 0 and 1 according to Table 1. Stages encoded as '0' are often referred to as 'Non-Aggressive', and the ones encodes as '1' as 'Aggressive'.

## 4.1 Data Pre-processing

Prior to any modelling work the dataset is normalised in order to eliminate the high variances between the gene's intensities (quantile normalisation). After normalisation, the student's distribution t-test is used as an initial gene-filter.

Based on the p-values the genes from the Sanchez-Carbayo dataset were reduced from the original 22243 genes down to a set of 500 genes.

## 4.2 Radial-Basis-Function Linguistic Modelling

The RBF-NF model was then developed as described in Section 2. The methodology was applied to the Sanchez-Carbayo Dataset, to reduce the number of genes. As previously discussed the Sanchez-Carbayo dataset consists of 90 patients and 22283 genes. A pre-selection of the genes was made using the top 500 genes as selected with the t-test. After this preliminary gene selection (pre-filtering) the entropy ($H_A$) (Eq, 2) is calculated based on the membership function ($\mu_i$). The median of the $\mu_i$ and the $H_A$ is calculated, a second threshold is established for both parameters (for $\mu_i$>.5 and for $H_A$<.4) then the Eq. 3 is applied. The first gene signature was developed with 250 genes. This number of genes was selected to compare the resulting modelling performance to the Sanchez-Carbayo results. The results are shown in Table 2.

The classification functions of Specificity, Sensitivity and Accuracy are used as measures of performance (Braga-Neto and Dougherty, 2004). The resulting model consisted of 10 rules only.

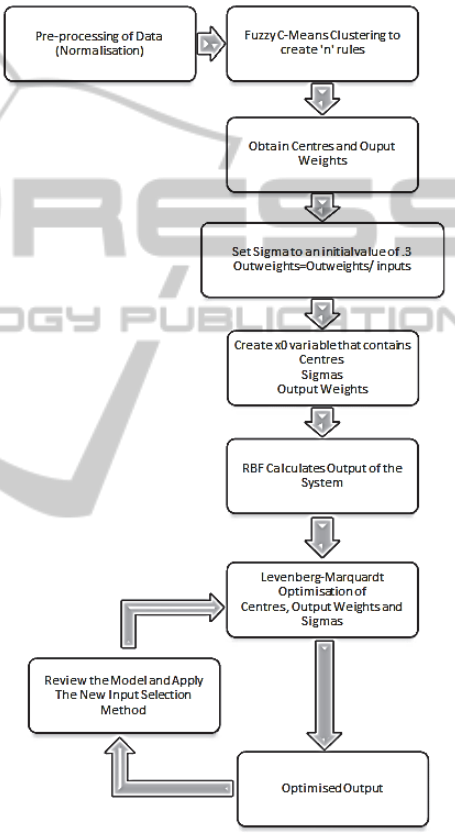


Figure 2: RBF-NF fuzzy entropy feature selection.

Table 2: RBF input selection 250 genes.

| | Training | Testing |
|---|---|---|
| Specificity | 100% | 100% |
| Sensitivity | 100% | 93% |
| Accuracy | 100% | 96% |

The data samples were randomly separated into 'Training' and 'Testing' datasets. The training set is only used to train the model. The testing dataset is only used after the model training is finished in order to test the generalisation performance of the

model (i.e. on 'unseen' by the model data), as a form of cross-validation.

The study from Sanchez-Carbayo on the same data-set used the popular (for microarray data analysis) method of support vector machines (SVM) to predict the tumour stage with 250 genes.

Table 3: Results Sanchez-Carbayo and RBF.

| | Sanchez-Carbayo | RBF Input Selection |
|---|---|---|
| Accuracy | 89% | 96% |

The results show that the two resulting models have a similar level of performance, using the exact same number of genes; however the RBF-NF method shows a slightly improved accuracy (+7%) as compared to the SVM model. On the second modelling attempt, in order to further examine the presented methodology, the RBF-NF model was compared to the Martin Lauss publication (Lauss et al., 2010). They used 201 genes for the prediction of the stage of cancer, based on a different dataset. Tables 4 and 5 summarise the RBF-NF modelling results and the Lauss results (SVM).

Table 4: 201 Genes – RBF-NF.

| | Training | Testing |
|---|---|---|
| Specificity | 90% | 100% |
| Sensitivity | 100% | 100% |
| Accuracy | 98% | 100% |

Table 5: Results comparison: Lauss vs. RBF-NF.

| | Martin Lauss | RBF Input Selection |
|---|---|---|
| Accuracy | 87% | 100% |

A third and final model was created, with just 10 genes, to investigate the generalisation performance of the methodology with a very low number of genres. Table 6 summarises the resulting model.

Table 6: RBF-NF input selection: 10 Genes.

| | Training | Testing |
|---|---|---|
| Specificity | 90% | 93% |
| Sensitivity | 96% | 93% |
| Accuracy | 95% | 93% |

As suggested by this table, a performance drop is observed, as compared to the 201 and 250 gene models, however one can say that the performance of the model did not decrease significantly and it still outperforms the previously developed more complex models presented in (Sanchez-Carbayo et al., 2006); (Lauss et al., 2010). Apart from the gene signature identification (10 genes) the modelling structure presented in this paper maintains a transparent Fuzzy Logic-type linguistic rule-base. Figure 3 shows a sample of the rule-base describing the behaviour of the model. For simplicity, just two rules are shown (one for 'low stage' and one for 'high stage') for five out of the 10 genes in the gene signature. Two of the linguistic IF-THEN rules that describe the model are shown below to demonstrate the transparency (interpretability) of the modelling method.

Rule 9:

IF: Gene RPS6 is Medium and
Gene PHB is Medium and
Gene LRP1 is Medium and
Gene CCND2 is Medium and
Gene SERP1 is Medium
THEN the Cancer Stage is Non-Aggressive

Rule 5:

IF: Gene RPS6 is Medium-High and
Gene PHB is Medium-High and
Gene LRP1 is Medium-High and
Gene CCND2 is Medium-High and
Gene SERP1 is Medium-High
THEN the Cancer Stage is Aggressive

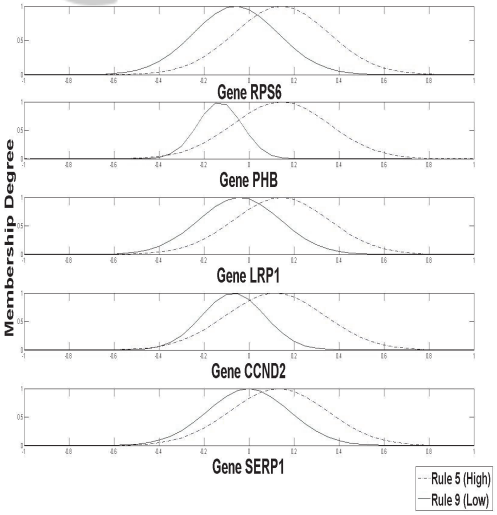


Figure 3: Neural-fuzzy rules.

Table 8 shows the 10 top ranked genes. The 10-gene signature has been confirmed from clinicians that is medically relevant; for example, genes RAPIB, CCND2 and SERP1 are known to be linked to bladder cancer. The corresponding numerical values of the linguistic hedges 'high', 'medium' etc. are determined by the optimisation algorithm via the

training data-set. The linguistic interpretation of the normalised gene intensity is shown in Table 7.

Table 7: Gene range.

| Gene Intensity | Range |
|---|---|
| Very Low | -1 to -0.72 |
| Low | -0.71 to -0.44 |
| Low Medium | -0.43 to -0.16 |
| Medium | -0.15 to 0.12 |
| Medium High | 0.13 to 0.4 |
| High | 0.5 to 0.68 |
| Very High | 0.69 to 1 |

Table 8: 10-Gene signature.

| Gene Symbol | Gene Title |
|---|---|
| RPL34 | ribosomal protein L34 |
| RPS6 | ribosomal protein S6 |
| PRKAR1A | protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1) |
| CSRP1 | cysteine and glycine-rich protein 1 |
| PHB | prohibitin |
| LRP1 | low density lipoprotein receptor-related protein 1 |
| DAZAP2 | DAZ associated protein 2 |
| RAP1B | RAP1B, member of RAS oncogene family |
| CCND2 | cyclin D2 |
| SERP1 | stress-associated endoplasmic reticulum protein 1 |

## 5 CONCLUSIONS

This paper introduces a feature selection algorithm based on Fuzzy entropy and a RBF Neural-Fuzzy structure that links directly the fuzzy entropy to the relative significance of the inputs of the model. This significance measure is used to rank the inputs of the model via an iterative algorithm. The proposed methodology has successfully been applied to the case study of bladder cancer prediction with respect to the 'stage' of the tumour. Compared to previous modelling attempts (Sanchez-Carbayo et al., 2006); (Lauss et al., 2010) based on SVM, the RBF-NF input selection method shows improved performance in the same datasets. The attractiveness of this method is on the transparency that the rule-base exhibits and the good generalisation performance (even with just 10 genes) as compared to previous modelling attempts on the same dataset (250 and 201 genes). The rule-base's transparency and interpretability, can aid the clinicians to directly interrogate the resulting model (human-centric system) and examine how the model uses individual genes and their intensity to provide predictions on the stage of bladder cancer. Further work should focus on predicting other cancer-related markers towards a more comprehensive predictive model. The biggest challenge though is presented in the generalisation ability of such data-driven models as identified by other research results too. Models that are trained based on a specific patient cohort should be tested against data from other cohorts to establish the developed models' generalisation performance and predictive robustness.

## REFERENCES


Al-Sharhan, S., F. Karray, et al. (2001). *Fuzzy entropy: A brief survey*, Melbourne.

Braga-Neto, U. M. and E. R. Dougherty (2004). "Is cross-validation valid for small-sample microarray classification?" *BIOINFORMATICS* 20(3): 374-380.

Burke, H. B., P. H. Goodman, et al. (1997). "Artificial neural networks improve the accuracy of cancer survival prediction." *Cancer* 79(4): 857-862.

Dunn, J. C. (1973). "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters." *Cybernetics and Systems* 3(3): 32-57.

Lauss, M., M. Ringnér, et al. (2010). "Prediction of stage, grade, and survival in bladder cancer using genome-wide expression data: a validation study." *Clinical Cancer Research* 16(17): 4421-4433.

Levenberg, K. (1944). "A method for the solution of certain non-linear problems in least squares." *The Quarterly of Applied Mathematics* 2(2): 164-168.

Pal, N. R., K. Aguan, et al. (2007). "Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering." *BMC Bioinformatics* 8(1): 5-5.

Panoutsos, G. and M. Mahfouf (2010). "A neural-fuzzy modelling framework based on granular computing: Concepts and applications." *Fuzzy Sets and Systems* 161(21): 2808-2830.

Sanchez-Carbayo, M., N. D. Socci, et al. (2006). "Defining molecular profiles of poor outcome in patients with invasive bladder cancer using oligonucleotide microarrays." J*ournal of Clinical Oncology* 24(5): 778-789.

Takagi, T. and M. Sugeno (1985). "Fuzzy identification of systems and its applications to modeling and control." *IEEE Transactions on Systems, Man and Cybernetics* 15(1): 116-132.