A fuzzy rule based approach to identify biomarkers for diagnostic classification of cancers

Nikhil R. Pal

***Abstract*—An important problem for doctors is to identify a small set of useful biomarkers (not all related genes) that can discriminate between different subgroups of cancers which appear similar in routine histology. Here we propose a method for simultaneous feature/gene selection and rule generation for the same problem. Since the feature selection method is integrated into the rule base tuning, it can account for possible subtle nonlinear interaction between features as well as that between features and the tool, and hence can identify a useful set of features for the task at hand. We applied our method to find biomarkers for a group of four childhood cancers that is collectively known as *small round blue cell tumors*. For this data set first we have used a neural network to reduce the dimension of the data and then applied our method to find biomarkers and rules. Our system could find only eight genes including a novel gene that can do the diagnostic prediction task with a high accuracy. The system can be extended to non-classification applications also.**

***Index Terms*— Identification of Biomarkers, Fuzzy rules, Cancer Subgroups, Gene Selection**

I. Introduction

The worldwide incidence rates of childhood cancers vary between 155 per million persons (in Nigeria) and 40 per million persons (in the Indian population of Fiji) [6], [8]. Between different childhood cancers, the group of small round blue cell tumors (SRBCTs) is the third most frequently occurring type (18%). The SRBCT group consists of neuroblastoma (NB, 7%), non-Hodgkin lymphoma (NHL,6%), rhabdomyosarcoma (RMS, 3%), and Ewing sarcoma (EWS, 2%) [6], [8]. These heterogeneous types of cancer present a similar histology of small blue tumor cell and thus often leads to misdiagnosis. Accurate diagnosis of these subgroups is important because the treatment options, monitoring of the responses and prognosis may vary widely between these subgroups. For problems like SRBCTs gene expression data are extensively used for categorization of cancers into different diagnostic subgroups [1], [2], [5]. This problem has been attempted to be solved by many researchers using different machine learning tools including neural networks, support vector machines, nearest shrunken centroid methods [1]- [5]. However, these methods cannot help doctors in understanding the interaction between genes that are identified as possible signature of a particular SRBCT type by a machine learning tool. For other types of cancers also many methods are proposed for identification of biomarkers [1], [9], [10], [24] and those methods also usually suffer from the same problem.

Nikhil R. Pal is with the Electronics and Communication Sciences Unit, Indian Statistical Institute, 203, B.T. Road, Calcutta-700108, (Email: nikhil@isical.ac.in)

Fuzzy rule based systems although are easily understandable by doctors, these have not been adequately explored to solve such problems in bioinformatics. Another very attractive attribute of fuzzy rule based systems for such critical applications is that fuzzy systems are not likely to make poor generalization because of its logical reasoning architecture. In this study we focus on classification of various groups of SRBCT childhood cancers. We intend to identify a small set of biomarkers to design *human readable* diagnostic prediction systems for classification of the four categories of SRBCTs.

Gene expressions profiles are usually of very high dimension. Extraction of useful rules in such a high dimension is very difficult. For the sake of argument, even if we assume that we can extract rules, for bioinformatic applications of this kind, such rules will not be much useful because human beings cannot interpret rules even in 20 dimension; the question of interpretability of rules does not make much sense when the dimensionality is several thousand. So there is an important problem of reducing the dimension before we can extract useful diagnostic rules. Moreover, for a given diagnostic classification problem, it is expected that a small set of genes will be adequate. So we face two tasks: finding a minimal set of genes that has an *adequate* discriminating power to categorize the subgroups and designing of a *human understandable* prediction system using the selected genes to classify unseen examples.

Most supervised methods for feature selection ignore the learning machine that is used to design the prediction system. Some methods, although, take into account the learning machines, they remove one (or a set of features) at a time in a stepwise manner. Such a method cannot capture the subtle nonlinear interactions that may exist between different genes/features and consequently, one ends up with more features than what is needed to solve the problem. Use of exhaustive search in conjunction with the machine learning tool that will be used to design the final diagnostic system, in principle, can solve the problem, but is computationally prohibitive.

In [22] we have proposed a method that uses neural networks and relational fuzzy clustering to discern a small set of genes with adequate discriminative power. These selected genes can be used to design useful classifiers using several machine learning tools.

In this paper we use a different approach for simultaneous feature/gene selection and fuzzy rule extraction. Our learning scheme can take into account the subtle interaction between genes and fuzzy rule based systems, and hence can pick up a small set of discriminative genes and extract useful rules for diagnostic prediction.

II. Simultaneous Feature Selection and Rule Extraction

Let $\mathbf{x} = (x_1, x_2, ..., x_p)^t \in R^p$ be an input vector and let there be c classes. Let $X = X_1 \cup X_2 \cup X_3$, $X_i \cap X_j = \Phi$, $i = 1, ..., c$ be the training data, X_i be the training data corresponding to class i.

Here we shall be considering classification rules of the following form : If x_1 is HIGH and .... and x_p is MEDIUM then class is k; $k \subset \{1, 2, ..., c\}$.

In general, R_i :If x_1 is $A_{1,i}$ and .... and x_p is $A_{p,i}$ then class is k; $k \in \{1, 2, ..., c\}$. Here $A_{j,i}$ is the ith fuzzy set (linguistic value) defined on the jth feature (linguistic variable). Let there be $N = \sum_{i=1}^{c} n_i$ rules, with n_i rules for class i.

The firing strength of a rule is computed using a T norm. Let α_i be the firing strength of the ith rule and $d = argmax_i\{\alpha_i\}$. Suppose the d-th rule R_d is associated with class k then x is classified as from class k.

Our objective here is to eliminate the features/genes that have poor discriminating power/ that can confuse the class boundaries / that are redundant. In [11] we proposed a neural network based method for simultaneous feature selection and system identification that through the process of learning modulates a feature value to zero, if the feature is bad or redundant. However, we cannot use such a concept here as zero membership value of any atomic antecedent clause will make the firing strength of the associated rule equals to zero as $T(0, \alpha)$=0. Here we like to modulate the membership values so that for a bad or redundant feature the modulated membership value from the associated antecedent clause becomes unity (1) because $T(1, \alpha) = \alpha$. If we can do so, then the associated membership value will not have any effect on the firing strength of the rule. In order to eliminate the effect of a bad/redundant feature completely, for every membership function defined on that feature the modulated membership value should be unity or very close to unity irrespective of the value of the linguistic variable. In [7] we attempted to solve this problem using the same basic principle with a 4-layer neuro-fuzzy system. Such a neuro-fuzzy system suffers from conflicting rules that need to be eliminated by some post-processing after the initial system is extracted. Moreover, even for moderately high dimensional data, the network size grows very rapidly and thereby limits its practical utility. The method proposed here eliminates all these problems.

In order to realize the proposed philosophy we use a modulator function : $M(\lambda) = exp(-\lambda^2)$. The modulated membership value is computed as $\mu' = \mu^{M(\lambda)}$. Here λ is a scalar variable and we call λ the modulator parameter which modulates the membership value, μ. To make it clear, let μ_j be a membership value associated with the jth membership function defined on the ith feature. Then the modulated membership value μ'_j is computed as $\mu'_j = (\mu'_j)^{(exp(-\lambda_i)^2)}$, Note that, for every linguistic variable there is *exactly one* modulator irrespective of the number of linguistic values defined on that variable.

Now if $\lambda_i \approx 0$ then $\mu \approx \mu'$. In other words, when $\lambda_i \approx 0$, positive or negative, the modulated membership values are nearly equal to original membership values. But if the magnitude of λ_i is very high, then $\mu \approx 1$, i.e., the modulated membership values are nearly 1 irrespective of the original membership values. Thus our objective is to set the modulator to a high value (high magnitude) if the feature is bad or redundant and to zero or nearly zero, if the feature is important for the discrimination task. We achieve it through learning using training data as explained next.

Before we describe the training process, a few other issues need to be discussed. There are many T norms. Product and minimum are the two most commonly used 'T' norms for computation of firing strengths. But use of product for classification problems is counterintuitive. Let us consider a rule :
If x_1 is A_1 *and* x_2 is A_2 *and* ... x_p is A_p then the class is k. Suppose for an input $\mathbf{x} = (x_1, x_2, ..., x_p)^T$ each x_i, $i = 1, 2, .., p$, has a membership of 0.9 in the respective fuzzy set A_i, $i = 1, 2, ..., p$. Thus, for this input if *product* is used as the operator for intersection then the firing strength of the rule will be $(0.9)^p$. Therefore, for a reasonably big p, the firing strength reduces almost to zero, though each of the input components has a high membership of 0.9 in the corresponding fuzzy set. Therefore, the use of product as an operator for intersection is not intuitively appealing. Note that, product can do a good job for function approximation type applications where the defuzzified output is computed as a convex combination of the peaks of the output fuzzy sets.

Although minimum is intuitively a very plausible intersection operator, it is not differentiable and hence we shall not be able to use it in our learning algorithm. So we use a softer but differentiable version of the minimum, $softmin$:

$$softmin(x_1, x_2, ..., x_p, q) = \left(\frac{x_1^q + x_2^q - ... + x_p^q}{p}\right)^{\frac{1}{q}}.$$

As $q \to -\infty$, $softmin$ tends to the minimum of all x_i's, $i = 1, 2, .., p$. Based on empirical exercises we found that $q = -11$ can realize a very good approximation to the minimum and that is what we use in all results reported. Thus, the firing strength α_i of the ith rule R_i is computed as :

$$\alpha_i = \left(\sum_{j=1}^{p} \frac{(\mu_{j,i})^{exp(-\lambda_j)^2)}}{p}\right)^{\frac{1}{q}}. \quad (1)$$

In (1), $\mu_{j,i}$ is the membership value of the jth feature to the ith the membership function (associated with the ith rule) defined on x_j. Given an initial set of rules, for identification of redundant/bad features we use gradient descent training to minimize the error function [19]:

$$E^X = \sum_{x \in X} E_x = \sum_{x \in X} (1 - \alpha_c + \alpha_{\neg c})^2. \quad (2)$$

In place of (2) the usual square error term can also be used. In (2), $x \in X$ is from the class c and R_c is the rule from class c giving the maximum firing strength α_c for x. Also $R_{\neg c}$ is the rule from the incorrect classes having the maximum firing strength $\alpha_{\neg c}$ for x.

We use the online version of the algorithm where for each input data, the modulators associated with the best and worst

rules are updated using the instantaneous error function E_x. Note that, if every rule has p antecedents then all modulators will be updated with every input. For notational clarity, unless it is needed, we shall drop the subscript x from E_x.

The modulator values are randomly initialized such that the modulated membership values for all linguistic values for all possible input values are practically 1. In other words, the training is started with modulator values such that all features pretend to be bad/redundant. Although, the algorithm works fine even when all modulators are initialized to the same high value, it is better to have some randomness in their initial choices so that to start with no feature is important but their initial modulators are slightly different.

The Modulator Learning Algorithm

The tuning process is repeated until E^X becomes negligible / the classification error is satisfactory / the maximum number of iterations is reached.

Begin

Choose: learning parameter, η_m; a parameter reduction factor, $0 < \varepsilon < 1$; maximum number of iterations, *maxiter*.

Compute the error E_0 and misclassification M_0 for the initial rule base R^0.

$t \leftarrow 1$

While ($t \leq maxiter$) do

For each vector $x \in X$

Find the rules R_s and R_{-s}.

Modify the parameters of rules and as follows.

For k=1 to p

$$\lambda_s^{new} = \lambda_s^{old} - \eta_m \frac{\partial E}{\partial \lambda_s^{old}}$$

End For

End For

Compute the error E_t^X for the new rule base R^t.

Compute the misclassification M_t for R^t.

If $M_t > M_{t-1}$ or $E_t^X > E_{t-1}^X$

then

$\eta_m \leftarrow (1-\varepsilon)\eta_m$

$R^t \leftarrow R^{t-1}$

/* Since the error is increased, we reduce the learning coefficient and restore the old modulator values . */

If $M_t^X = 0$ or $E_t^X \approx 0$

then Stop

$t \leftarrow t+1$

End while

End

Note that in the above algorithm, we do not tune the center and spread of the associated membership functions. One may be tempted to do so adding two more sets of update equations for the centers and spreads in the while loop:

$$v_{cs}^{new} = v_{cs}^{old} - \eta_m \frac{\partial E}{\partial v_{ck}^{old}}$$

$$v_{\neg ck}^{new} = v_{\neg cs}^{old} - \eta_m \frac{\partial E}{\partial v_{\neg ck}^{old}}$$

$$\sigma_{ck}^{new} = \sigma_{cs}^{old} - \eta_s \frac{\partial E}{\partial \sigma_{ck}^{old}}$$

$$\sigma_{\neg ck}^{new} = \sigma_{\neg cs}^{old} - \eta_s \frac{\partial E}{\partial \sigma_{\neg ck}^{old}}$$

In the above equations η_m and η_s are the learning coefficients. But simultaneous tuning of centers and spreads of membership functions with feature modulators is not recommended. If membership parameters are to be tuned, then that must be done in a separate phase after the learning of modulators is over and the rule base is simplified. The membership parameter tuning looks at the performance of each rule, while the modulator tuning looks at the global picture taking all rules (all membership functions defined on a linguistic variable) together. Consequently, the tuning algorithm may become unstable, if we try to tune all parameters simultaneously.

III. Finding an Initial Rule Base

The above modulator learning algorithm requires an initial rule base. We use the fuzzy c-means (FCM) [21] clustering algorithm to cluster each $X_i \subseteq R^p$, the training data from the ith class, into say n_i clusters. Each cluster is expected to represent a dense/important area in the input space which is quantized by the associated cluster centroid. Each such cluster is converted into a fuzzy rule of the form:

R_s : if x is CLOSE TO v_i then the class is k.

Here v_i is the centroid of the ith cluster. The fuzzy set "CLOSE TO" can be modeled by a multidimensional membership function such as

$$\mu_{CLOSETO} \quad \mathbf{v}_i(\mathbf{x}) = \exp^{-\frac{\|x-v_i\|^2}{\sigma_i^2}},$$

where $\sigma_i > 0$ is a constant. This is equivalent to using prototypes with hyperspherical zones of influence centered at $\mathbf{v}_i$. Such a classifier may not perform quite well when different features have considerably different variances. Moreover, for a better readability "x is CLOSE TO $\mathbf{v}_i$" is written as a conjunction of p atomic clauses :

x_1 is CLOSE TO v_{i1} AND $\cdots$ AND x_p is CLOSE TO v_{ip}.

Thus, the i-th rule R_i representing one of the c classes takes the form R_i : if x_1 is CLOSE TO v_{i1} AND $\cdots$ AND x_p is CLOSE TO v_{ip} then class is k. We emphasize that the form with the multidimensional membership and with atomic antecedent clauses are not necessarily the same.

In this investigation, the fuzzy set CLOSE TO v_{ij} is modeled by a Gaussian membership function :

$$\mu_{ij}(x_j; v_{ij}, \sigma_{ij}) = \exp(-(x_j - v_{ij})^2/{\sigma_{ij}}^2).$$

The center of the Gaussian membership function is initialized with the corresponding component of the cluster center. For the membership functions defined on the jth feature the spreads can be initialized with the standard deviation of the j-th component of the training data included in the associated cluster as defined by the maximum membership hardening rule.

IV. Results

A. Data Set.

In [5] Khan et al. used 2308 genes that passed a filter satisfying a minimum expression level. We also use the same set of 2308 genes. There are 88 samples of which 63 (EWS:23, NHL:8, NB:12, and RMS:20) samples are used for training. The remaining 25 includes five samples which are later detected to be of non-SRBCT types. Hence for blind testing of the system we use 20 (EWS:6, NHL:3, NB:6, and RMS:5) samples. Other investigators also have used this training-test partition. The data set is available at http://research.nhgri.nih.gov/microarray/Supplement/.

B. Initial dimensionality reduction with neural networks

The dimensionality of this data set is too high to use a clustering algorithm to extract a useful initial rule base. We first use a neural network to select an initial set of features. Fuzzy clustering is done on this reduced data set. Such neural networks have also been used to select genes in [22]. We use a modified multilayered perceptron (MLP) network [11] with online feature selection capability. This network is called a Feature Selection MLP (FSMLP). Conceptually, each input node (hence each gene) of the FSMLP has a gate associated with it. At the beginning of training, these gates are kept almost closed, and the training algorithm opens the required gates (allows features to enter the network) depending on the ability of features to reduce the training error. The same set of twenty genes selected by FSMLP in [22] are used here. These genes are listed in Table 1. Note that these 20 genes have many redundant genes as reported in [22].

C. Final gene selection and fuzzy rule extraction

Now the training data from each class is divided into two clusters using FCM. This gives us eight rules, two rules representing each of the four categories of SRBCTs. The desired number of clusters/rules for each class may be decided using cluster validity indices as done by many authors [4]. However, we do not use any cluster validity indices here for two reasons: (1) Our objective is to show the feature selection ability using fuzzy rule base and (2) use of cluster validity for rule extraction is conceptually not an appealing approach because a cluster representing a rule may not necessarily be a cluster in the pattern recognition sense. Moreover, the training data may not have any cluster in the pattern recognition sense yet rules can be extracted using clustering [3].

The rules in the initial rule base use all twenty genes listed in Table I. This rule base makes only one training error. We analyzed the rule usage on the training set and found that two rules did not take part in decision making. So we deleted those two rules. The reduced rule base is now tuned for gene selection. At the end of the training, we find that 12 features are not important. The rule base is then simplified keeping only the eight genes/features. These selected genes are marked by asterisks in Table I. Table I also shows the modulator values after training. The rule base is now simplified removing the antecedent clauses relating to bad/redundant genes. The rule base may now be further tuned to refine its centroids to adapt itself to its new environment ignoring the modulators, i.e., setting all modulator values to zeros (no modulation). Although we could have used better methods, as discussed earlier, for initialization of the spreads of the membership functions, we have used 0.5 (also experimented with a few other choices) for all membership functions. Since the spread of every membership function is very low, the specificity of each rule is very high which is very important for critical applications. This rule base with six rules leads to one training error and zero test error.

D. Relevance of the identified Biomarkers in Cancer Biology

Figure 1 shows a scatterplot of four of the eight genes each of which exhibits a strong class specific signature for one of the four cancer types. These eight genes identified by our system seem to be involved in the biological process of cancer. For example, this set of eight genes includes an interesting gene EH domain containing 1 (EHD1) which was not found as important by Khan et al. [5] and Fu & Fu-Liu [10]. As reported in [22] EHD1 is upregulated in metastatic colon cancer compared to colon tumors (GEO: GPL96 208112). On the other hand, according to GEO profiles, EHD1 is found to be downregulated in EWS (GEO: GPL1977 1465) and in B-cell lymphoma (GEO: GPL176 5453). For the SRBCT cases, we have found that EHD1 upregulated in Non-Hodgkin Lymphoma (NHL) along with a few cases of RMS but it is moderately expressed for most of the of RMS and EWS samples.

Like other investigators [5], [22], we also found fibroblast growth factor receptor 4 (FGFR4) as very important with distinct class-specific signature, see Fig. 1. Figure 1 reveals that it is highly expressed for the RMS group but for the other three subgroups this gene is practically unexpressed. This gene, thus, has a very strong RMS-specific signature. This gene has been found to play an important role in cancer biology. For example, overexpression of FGFR4 has been found in different cancers [13] [15]. On the other hand, in lung adenoarcinoma, FGFR4 is found to be downregulated [16].

Panel (b) in Figure 1 clearly reveals that FCGRT (Fc fragment of IgG, receptor) has a distinct EWS specific signature because it is moderate to highly expressed for the EWS group and is downregulated for the other three groups. This gene plays significant roles in other types of cancers too. For example, in [23] authors identified a list of 26 prognostic genes offering predictive information on patient survival for lung cancer. They found that a higher expression level of FCGRT corresponds to better survival outcome.

We found that AF1Q is moderate to highly expressed for the neuroblastoma group and is downregulated for the other three groups of tumors. Other researchers have also found this gene to be important in cancer biology [17], [18].

Panel (c) in Figure 1 depicts that for the present case, CDH2 is underexpressed in the EWS and NHL groups of tumors, while for the NB class it is moderate to highly upregulated. For a few RMS cases also it is found to be moderately expressed.

Authors in [22] reported that according to GEO profiles CDH2 is found to be downregulated in EWS which is consistent with our finding in this study. Usually the low expression of CDH2 leads to tumor invasiveness [12].

The gene PMS2L12 is moderate to highly expressed for the neuroblastoma group and is practically unexpressed for the NHL group of tumors; while for the EWS group for one sample it is very highly expressed.

As mentioned earlier, many other investigators analyzed this data set with a view to find biomarkers [2], [5], [20]. Here are some of the results from the literature. Khan et al. [5] used PCA and then used a single layer neural network. They suggested a list of 96 genes as important for these four groups of tumors. The nearest centroid method with shrunken centroids is used by Tibshirani et al. [2]. This method shrinks the centroid of each class towards the overall centroid using the within-class standard deviation of each gene. The shrunken centroid method suggested 43 genes as important.

Fu and Fu-Liu [10] applied the method of Ramaswamy et al. [9] on the SRBCT data set to select a set of 8 genes yielding zero training error but 90% test accuracy. This is a method based on support vector machine (SVM) [9]. Fu and Fu-Liu [10] also proposed a method using SVM for gene selection. For the SRBCT data set, they found nineteen (19) genes as important. These genes can achieve 100% accuracy both on the training and test data sets. In [22] seven genes are found and this set contains four of the eight genes found here, but does not include the genes HCLS1 and PMS2L12 identified by the fuzzy system. Moreover, the gene PMS2L12 found by this method is not considered important in [2], [5].

TABLE I

THE LIST OF TWENTY BEST GENES SELECTED BY FSMLP. THE EIGHT GENES SELECTED BY THE FUZZY RULE BASE ARE INDICATED BY ASTERISKS

Gene ID	Name	modulator values
FGFR4 (*)	fibroblast growth factor receptor 4	0.0
EST	EST	1.0947
FCGRT (*)	Fc fragment of IgG, receptor, transporter, alpha	0.0
AF1Q (*)	transmembrane protein	0.0003
HCLS1 (*)	hematopoietic cell-specific Lyn substrate 1	0.0117
NAB2	NGFI-A binding protein 2 (EGR1 binding protein 2)	0.1823
CDH2 (*)	cadherin 2, N cadherin (neuronal)	0.0133
EHD1 (*)	EH domain containing 1	0.1454
HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	0.6583
LGALS3BP	lectin, galactoside binding, soluble, 3 binding protein (galectin 6 binding protein)	1.1903
BAT3	HLA-B associated transcript-3	0.7286
SGCA	sarcoglycan, alpha (50kD dystrophin-associated glycoprotein)	0.9867
ESTs (*)	ESTs	0.0
NOE1	olfactomedinrelated ER localized protein	0.4851
LSP1	lymphocyte-specific protein 1	0.9611
IFG2	insulin-like growth factor 2 (somatomedin A)	1.0822
PMS2L12 (*)	postmeiotic segregation increased 2-like 12	0.0474
NA	NA	1.1819
FVT1	follicular lymphoma variant translocation 1	0.2095
CCNB1	cyclin B1	1.1696

V. CONCLUSION

We have proposed a method for simultaneous feature selection and fuzzy rule extraction. Unlike other feature selection methods used in connection with fuzzy rules, our method can account for interaction between features as well as that between features and the tool, and hence can find a small set of important features. The method is applied to identify a small set of biomarkers for diagnostic prediction of the SRBCT group of childhood cancers. For this data set we first reduced the dimension using a neural network based method and then applied our method. Our method could find a set of eight genes for the task. This set of genes includes a new gene that others have not found as important.

Note that, depending on the learning parameters and the initial rule base different trials of the algorithm may find different sets of important features because gradient descent search usually settles at a local minima. If that happens, then that will indicate existence of different subsets of genes that can do the task well. To decide on the number of useful features, we need to choose a threshold for the modulator values. Further investigation is needed to come up with useful guidelines for this. Also investigations need to be done to analyze the performance of the algorithm when the dimension of the input is very high, say a few thousand. The method can be extended to function approximation type applications with both Mamdani type and Takagi-Sugeno type modeling.

REFERENCES

[1] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**: 531-537.

[2] Tibshirani R, Hastie T, Narasimhan B , Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS* **99**: 6567-6572.

[3] K. Pal, R. Mudi and N. R. Pal, A new scheme for fuzzy rule based system identification and its application to self tuning fuzzy controllers, *IEEE Trans. Systems Man and Cybern - B*, Volume 32, No. 4, pp 470-482, 2002.

[4] N. R. Pal, R. Mudi, K. Pal, and D. Patranabish, Rule extraction to exploratory data analysis for self-tuning fuzzy controllers, *Int. Jour. Fuzzy Systems*, Vol 6, No. 2, pp 71-80, 2004

[5] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **7**: 673-679.

[6] Kupfer GM (2003) Childhood Cancer, Epidemiology. *eMedicine* **[http://www.emedicine.com/ped/topic2585.htm]**

[7] D. Chakraborty and N. R. Pal, A neuro-fuzzy scheme for simultaneous feature selection and fuzzy rule based classification, *IEEE Trans. Neural Networks*, Vol. 15, No. 1, pp 110-123, 2004.

[8] Parkin DM, Kramarova E, Draper GJ, Masuyer E, Michaelis J, Neglia J, Qureshi S, Stiller CA (1999) International Incidence of Childhood Cancer, Volume II. *IARC Scientific Publications*.

[9] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, PoggioDagger T, Gerald W, Lodadagger M, Lander ES, Golub TR (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS* **98**: 15149-15154.

[10] Fu L. M. & Fu-Liu C. S. (2005) Evaluation of gene importance in microarray data based upon probability of selection. *BMC Bioinformatics* **6**: 67-78.

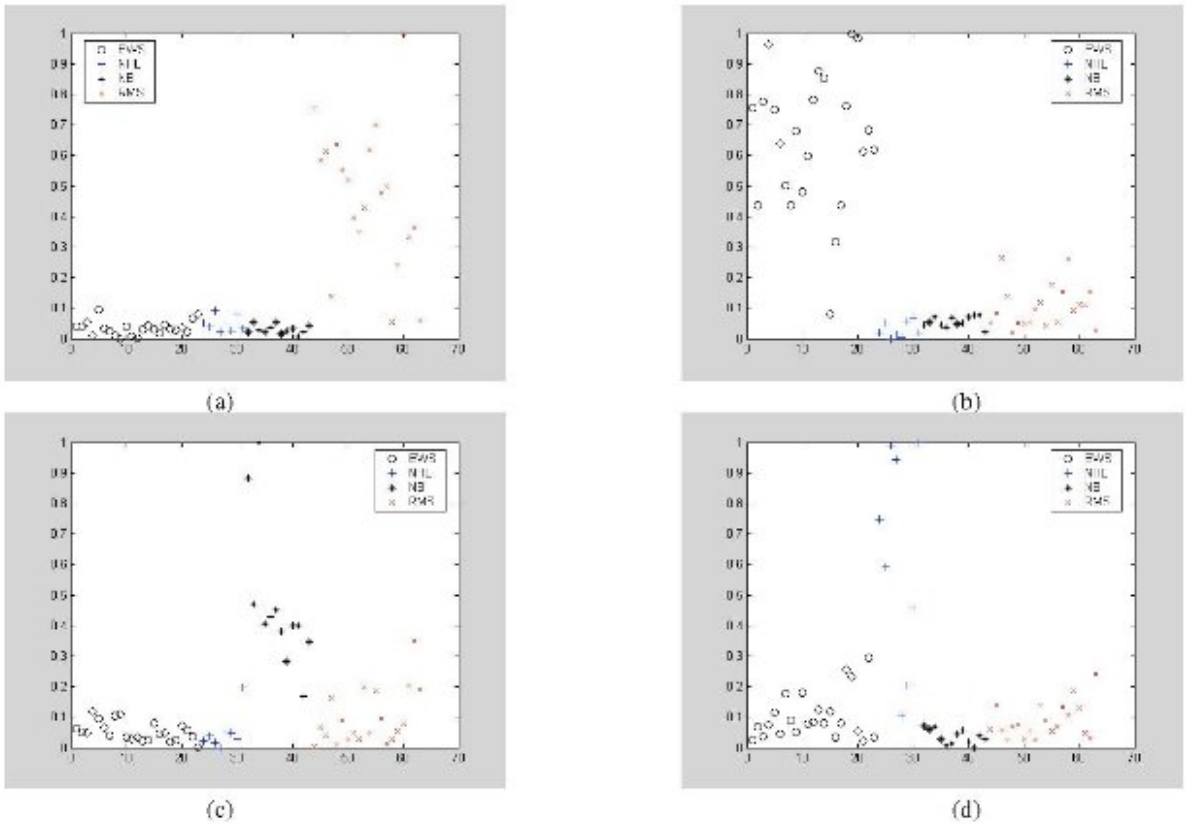

Fig. 1. Scatterplot of four identified genes in the training data: (a) FGFR4, (b) FCGRT, (c) CDH2 and (d) HCLS1

[11] Pal NR, Chintalapudi KK (1997) A connectionist system for feature selection. *Neural, Parallel and Scientific Computations* 5: 359-382.

[12] Hirohashi S, Kanai Y (2003) Cell adhesion system and human cancer morphogenesis. *Cancer Sci* 94:575-581.

[13] Qian ZR, Sano T, Asa SL, Yamada S, Horiguchi H, Tashiro T, Li CC, Hirokawa M, Kovacs K, Ezzat S (2004) Cytoplasmic expression of fibroblast growth factor receptor-4 in human pituitary adenomas: relation to tumor type, size, proliferation, and invasiveness. *J Clin Endocrinol Metab* 89:1904-1911.

[14] Wang J, Stockton DW, Ittmann M (2004) The fibroblast growth factor receptor-4 Arg388 allele is associated with prostate cancer initiation and progression. *Clin Cancer Res* 10:6169-6178.

[15] Ezzat S, Huang P, Dackiw A, Asa SL (2005) Dual inhibition of RET and FGFR4 restrains medullary thyroid cancer cell growth. *Clin Cancer Res* 11:1336-1341.

[16] Nakamura N, Iijima T, Mase K, Furuya S, Kano J, Morishita Y, Noguchi M (2004) Phenotypic differences of proliferating fibroblasts in the stroma of lung adenocarcinoma and normal bronchus tissue. *Cancer Sci* 95:226-232.

[17] Ise W, Zhu W, Chen HS, Cohen A (1995) A novel gene, AF1Q, fused to MLL in t(1;11)(q21;q23), is specifically expressed in leukemic and immature hematopoietic cells. *Blood* 85: 650-656.

[18] Ise W, Meshinchi S, Alonzo TA, Stirewalt DL, Robert B. Gerbing, Woods WG, Appelbaum FR, Radich JP (2004) Elevated expression of the AF1Q gene, an MLL fusion partner, is an independent adverse prognostic factor in pediatric acute myeloid leukemia. *Blood* 104: 3058-3063.

[19] S. L. Chiu, Method and software for extracting fuzzy classification rules bysubtractive clustering, *Proceedings NAFIPS 1996*., Berkeley, USA, pp 461-465, 1996.

[20] Lee, J. W. , Lee, J. B., Park, M. & Song, S. H. (2005) An extensive comparison of recent classification tools applied to microarray data, *Computational Statistics & Data Analysis*,48, 869-885.

[21] Bezdek, J. C., Keller J., Krishnapuram, R., Pal, N. R. (1999) **Fuzzy Models and Algorithms for Pattern Recognition and Image Processing**, *Kluwer Academic Publisher*

[22] Pal N. R., K. Aguan, A. Sharma and S-I Amari, Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering, *BMC Bioinformatics*, 8:5, 2007.

[23] J. S. Morris, G. Yin, K. Baggerly, C. Wu, and L. Zhang, Pooling information across different studies and oligonucleotide chip types to identify prognostic genes for lung cancer, *Methods of Microarray Data Analysis IV*. New York: Springer, 2005

[24] N. R. Pal, A. Sharma, S. Sanadhya and Kamneshu, Identifying Marker Genes from Gene Expression Data in a Neural Framework through Online Feature Analysis, *International Journal of Intelligent Systems*, Volume 21, Issue 4, April 2006, Pp: 453-467