Eng Int Syst (2015) 2: 97–112
© 2015 CRL Publishing Ltd

Engineering
Intelligent Systems

Type-2 fuzzy logic in breast cancer relapse time prediction using genetic algorithm

Hamid Mahmoodian

Department of Electrical Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Esfahan, Iran.
Phone: 0098-331-2291111, Fax: 0098-331-2291016, E-mail: h_mahmoodian@pel.iaun.ac.ir

Abstract-Microarray analysis and gene expression profile have been widely used in tumor classification, survival analysis and ER (Estrogen Receptor) status of breast cancer. Sample discrimination as well as identification of significant genes, has been the focus of most previous studies. Fuzzy Partitioning Method (FPM) has been introduced in our previous study which generates a rule base for fuzzy systems based on multi-level partitioning of gene expression profiles. The aim of this research is to improve proposed method by optimizing the size of selected genes in a procedure which has been statistically analyzed in this study. To handle the existence of uncertainties in proposed model, the TSK (Takagi-Sugeno-Kang) model of Interval Type-2 Fuzzy Logic System (IT2-FLS) has been used to predict the relapse time of breast cancer. In addition, a new method has been developed to consider the uncertainties of the fuzzy rules generated in the model. The results with 95% confidence interval show an improvement in the relapse time prediction with respect to the previous works.

Keywords: gene expression profile, breast cancer, fuzzy systems, relapse time, genetic algorithm

1. INTRODUCTION

Recently, Gene selection, dimension reduction of gene expression profiles and microarray analysis have been widely applied in diagnosis, survival analysis and or tumor classification of breast cancer [1-7]. The contribution of most of them is to discriminate the high and low risk cancerous samples [8-11]. Survival analysis or relapse time prediction of breast cancer has been also considered in some articles. For example, in [12] and [13], fuzzy TSK model and support vector regression have been used to predict the relapse time of breast cancer based on van't Veer dataset. In the former, correlation coefficient between relapse time and gene expression values have been used to generate antecedent parts of TSK model and in the latter, the predictor was based on support vector regression. Both of these studies try to predict the relapse time of breast cancer on a continuous scale similar to the Gruvberger's research for predicting the ER statues [14].

In our previous study [12], the ability of proposed algorithm (Fuzzy Partitioning Method- FPM) in generating rule base of fuzzy predictor was shown. However, there was no statistical analysis on the validity of the results. The aim of this study is to apply the FPM to a comprehensive procedure which includes optimization of the number of genes, different types of validation (e.g. cross validation and bootstrap) and some statistical analyses.

Different methods have been illustrated in the literatures for generating a rule base in a fuzzy system [15-18]. Support vector machine (SVM) has been also applied to make the rule bases based on the projection of the support vectors on the fuzzy space [19-20].

To train and test the proposed model two datasets proposed by van't Veer [1] (about 5000 genes after preprocessing with 77 and 19 samples as training and test samples respectively)

and van de Vijver [21] (36 of 234 samples that are lymph node negative) have been used in this study. The Relative Error (RE) and Square Correlation Coefficient (SCC) has been used as two criteria to analyze the results of prediction.

2. GENE RANKING

SNR (signal to noise ratio), PCC (Pearson correlation coefficient) and SVM-RFE [22] (support vector machine-recursive feature elimination) methods have been applied to sort the genes. SNR and PCC are known as univariate filter methods which sort the genes based on their discrimination abilities. The interactions between the genes are not considered and the gene selection procedure is completely independent of the classification procedure. The relapse time of tumors has been used to compute the correlation coefficients.

SVM-RFE, on the other hand, is a multivariate wrapper gene selection method which is embedded with SVM classifier and filters out low weighted genes recursively.

3. REVIEW ON INTERVAL TYPE-2 FUZZY LOGIC

3.1 Basic concepts and aggregation

The i^{th} ($i = 1, 2, \ldots, R$) rule of a typical IT2-TSK model is as follow [23-24]:

$$\text{if } x_1 \text{ is } A_1^k \text{ and } \ldots x_f \text{ is } A_f^k \text{ then } Z^k = \tilde{p}_0^k + \tilde{p}_1^k x_1 + \ldots + \tilde{p}_f^k x_f \tag{1}$$

where A_j^i and x_j are j^{th} ($j = 1, 2, \ldots, f$) IT2 membership function and j^{th} input respectively, output coefficients $p_{\overline{k}}^i \left(\overline{k} = 0, 1, \ldots f\right)$ are IT2 factors in range $[\underline{a}_{\overline{k}}^i, \overline{a}_{\overline{k}}^i]$ to handle the uncertainty in the outputs. Final output is equal to:

$$Z = \frac{Z_l + Z_r}{2} \text{ where } Z_l = \min_{\substack{\forall Z \\ \forall \mu}} \frac{\sum_{i=1}^{R} Z_i \mu_i}{\sum_{i=1}^{R} \mu_i}$$

$$\text{and } Z_r = \max_{\substack{\forall Z \\ \forall \mu}} \frac{\sum_{i=1}^{R} Z_i \mu_i}{\sum_{i=1}^{R} \mu_i} \tag{2}$$

Karnik-Mendel algorithm [25-27] should be applied to find $Y_l$and Y_r in (5). In above equation, we have $\mu_i \in \left[\underline{\mu}_i, \overline{\mu}_i\right]$, $Z_i \in \left[\underline{Z}_i, \overline{Z}\right]$ and $\underline{\mu}_i = \min(\overline{a}_1^i(x_1), \overline{a}_2^i(x_2), \ldots, \overline{a}_n^i(x_n)))$ where $\overline{A}_j^i\left(x_j\right)$ and $\overline{a}_j^i\left(x_j\right)$ are minimum and maximum range in FOU of j^{th} membership function in i^{th} rule of IT2-TSK model.

3.2 Definition of membership functions and their uncertainties

In this study, seven fuzzy sets such as figure 1 have been used to fuzzify the gene expression values. Hierarchical clustering method has been applied to cluster the gene expression values of each gene of training samples in three clusters (large, medium and small) for both positive and negative amounts. The mean of the amounts in the clusters is placed as vertices of the triangular fuzzy sets.

To measure the uncertainties for the linguistic terms, training samples were divided into 10 folds randomly and the negative and positive expression values were clustered into three separate clusters (total of 6 clusters) while one fold of samples has been left out. Let $\phi_{f,lv}^{\rho}$ be the mean of the gene expression values of lv^{th} ($lv = 1, 2, \ldots, 6$) cluster of f^{th} genes when ρ^{th} ($\rho = 1, 2, \ldots, 10$) fold is left out. Then the uncertainty of the lv^{th} fuzzy set (FOU) of f^{th} genes is considered by (3).

$$u_{f,lv} = max(\left|\phi_{f,lv}^{\rho_1} - \phi_{f,lv}^{\rho_2}\right| | \rho_1, \rho_2 = 1, 2, \ldots 10). \tag{3}$$

3.3 Rule mining

SVM has been considered to generate the rules in previous works [19-20]. In this study, the support vectors of the SVM classifier are projected on the space of fuzzy set to generate the antecedent parts of the fuzzy rules such as proposed method in [20].The consequent parts of the rules have been changed as TSK model such as equation (1). A two dimensional example is depicted in figure 2.

3.4 Rule pruning and fuzzy confidence

Similar to the proposed method in [12], the mean of the compatibility values for each rule (generated by support vectors) has been used in rule pruning and computing the fuzzy confidence values as well. Let $\mu_{A(\cdot)}^k$ be the membership degrees of the fuzzy item sets firing by a sample in k^{th} rule. Therefore, the amount of $\prod_{A(\cdot)} \mu_{A(\cdot)}^k$ can be considered as the compatibility between the samples and k^{th} rule. The mean of compatibility values can be calculated by (4) and (5) where $\overline{s}_0$ and $\overline{s}_1$ are the number of samples in two classes.

$$\text{comp}_{c0}\left(k^{\text{th}} \text{ rule}\right) = \frac{\sum_{\text{all samples of class0}} \prod_{A(\cdot)} \mu_{A(\cdot)}^k}{\overline{s}_0} \tag{4}$$

$$\text{comp}_{c1}\left(k^{\text{th}} \text{ rule}\right) = \frac{\sum_{\text{all samples of class1}} \prod_{A(\cdot)} \mu_{A(\cdot)}^k}{\overline{s}_1} \tag{5}$$

If the k^{th} rule is an extracted rule from samples of class 0 and comp_{c0} is greater than comp_{c1}, then the rule is considered in the rule base, otherwise it is pruned. In this case, the confidence value of the rule can be defined as:

$$FC\left(k^{\text{th}} \text{ rule}\right) = \text{comp}_{c0}\left(k^{\text{th}} \text{ rule}\right) - \text{comp}_{c1}\left(k^{\text{th}} \text{ rule}\right) \tag{6}$$

3.5 Uncertainty in the rules

To consider the confidence values in the TSK rules, a virtual feature $x_{\text{virtual}} = 1$ with upper and lower membership functions defined by: $\overline{w}_{\text{virtual}}^k = FC(k^{\text{th}} \text{ rule}) + u_{\text{virtual}}^k$ or $\underline{w}_{\text{virtual}}^k = FC(k^{\text{th}} \text{ rule}) - u_{\text{virtual}}^k (k = 1, 2, \ldots, R)$, has been added to the antecedent parts of the rules (u_{virtual}^k is uncertainty of the confidence values). By this definition, upper and lower membership degrees to the rules are calculated

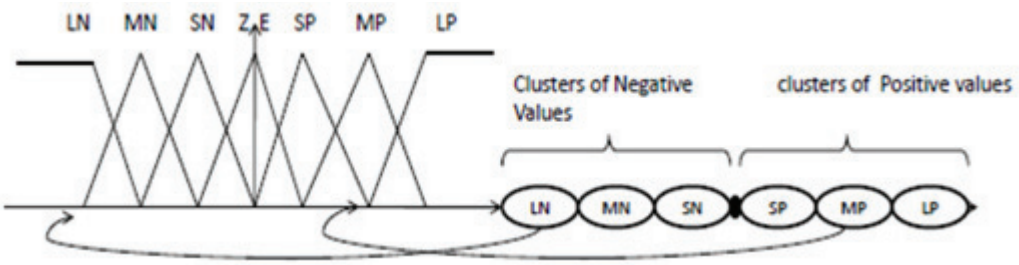

Figure 1 Seven fuzzy sets as fuzzifier.

Extracted rules are:
If x_1 is ZE and x_2 is SP then class A
If x_1 is SP and x_2 is SN then class B

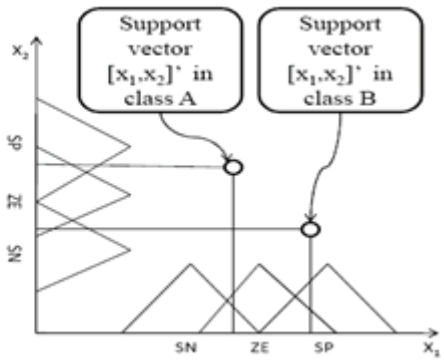

Figure 2 Two dimensional examples for fuzzy rule mining based on SVM.

by (7) and (8) respectively.

$$\overline{\mu}^k = \overline{w}_1^k \oplus \overline{w}_2^k \oplus \cdots \overline{w}_f^k \times \overline{w}_{\text{virtual}}^k \tag{7}$$

$$\underline{\mu}^k = \underline{w}_1^k \oplus \underline{w}_2^k \oplus \cdots \underline{w}_f^k \times \underline{w}_{\text{virtual}}^k \tag{8}$$

where '$\times$' is product aggregation operator and $\oplus$ is defined as arithmetic mean of the membership values in this paper. Genetic algorithm (GA) has been used to recognize u_{virtual}^k $(k = 1, 2, \ldots, R)$ for all rules which is explained in section 5.

3.6 Calculation of consequent parameters

The coefficients in consequent parts of the rules $(\tilde{p}^k)$ have been determined based on the proposed method in [30] to minimize the equation (9) which is a linear combination of two equations in (10) and (11).

$$J_C = \beta_{\text{Global}} J_{\text{Global}} + \beta_{\text{Local}} J_{\text{local}} \tag{9}$$

Such that $\beta_{\text{Global}} + \beta_{\text{Local}} = 1$, J_{Global} and J_{local} are the following objective functions:

$$J_{\text{Global}} = \sum_{j=1}^{s} [RRT\,(j) - Z\,(j)]^2 \tag{10}$$

$$J_{\text{local}} = \sum_{k=1}^{R} \sum_{j=1}^{s} \mu_k\,(j)\,[RRT\,(j) - Z\,(j)]^2 \tag{11}$$

where $RRT\,(j)$ is the real relapse time of the j^{th} training sample, $Z\,(j)$ is the predicted value by the model and $\mu_k\,(j)$ is the firing strength of k^{th} rule respect to the j^{th} training sample. The results of fuzzy model have been analyzed in two cases: 1) $\beta_{\text{Global}} = \beta_{\text{Local}} = 0.5$ and 2) $\beta_{\text{Global}} = 1$ and $\beta_{\text{Local}} = 0$.

4. FUZZY PARTITIONING METHOD (FPM)

FPM has been introduced in our previous work [12]. It includes two sub-algorithms called separation and final algorithms. Although FPM is explained in detail in [12] but it could be summarized as follow:

Let $\overline{G} \in \mathbb{R}^{f \times s}$ be gene expression profile with elements g_{ij} where it shows the expression value of the i^{th} gene $(i = 1, 2, \ldots, f)$ in the j^{th} sample $(j = 1, 2, \ldots, s)$ and suppose that $V \in \mathbb{R}^s$ is the vector of relapse time values and p is the number of levels in the rule mining procedure. Separation algorithm is starting by sorting the rows of $\overline{G}$ based on the absolute value of Pearson Correlation Coefficient (PCC) between the gene expression values and vector V. The new matrix will be called $G \in \mathbb{R}^{f \times s}$ and suppose that $\tilde{G} \in \mathbb{R}^{\tilde{f} \times \tilde{s}}$ is a subset of $G \in \mathbb{R}^{f \times s}$ which is called a 'group' in the algorithm. The separation algorithm includes the following steps:

- Set l equal to the number of the samples which should be added in the iterations (since the algorithm is based on the SVM binary classification, it is reasonable to select the $l$at least equal 2).
- Set $k = 1$
- Divide the samples of the group into two subgroups with size $k \times l$ and $\tilde{s} - k \times l$.
- Train the SVM classifier based on the samples of two subgroups of step 3. In this step, the two class labels 0 and 1 should be considered for the samples in two subgroups.

- Recognize the support vectors and construct the fuzzy rules after projection procedure.
- Prune the rules based on the explained rule pruning method.
- Calculate the confidence value of the rules.
- Set h^k as sum of all confidence values of the extracted rules in iteration k.
- If $(\tilde{s} - k \times l) > l$ then $k = k + 1$ and go to step 3, otherwise go to step 10.
- Find k which has maximum (h^k) and call it k^*.
- Divide the original group into two subgroups with size $k^* \times l$ and $\tilde{s} - k^* \times l$.
- For each subgroup, select the rule with the highest confidence value as the representative of the extracted rules.

The final algorithm is based on the partitioning the rows of the matrix $G \in \mathbb{R}^{f \times s}$ into the p levels and includes the following steps:

- Set ind_{level}=1 (ind_{level} points to the level number, e.g. first level)
- Set $ind_{\text{group}} = 2^{ind_{\text{level}}}$ (ind_{group} points to the number of group)
- Use separation algorithm to divide each group of the previous level into two subgroups
- Find the best rule of each group (call it R_{xw} where $x = 1, \ldots, ind_{\text{level}}$ and $w = 1, \ldots, ind_{\text{group}}$)

$$ind_{\text{level}} = ind_{\text{level}} + 1$$

- If $ind_{\text{level}} \leq p$ go to step2 otherwise go to step7
- Merge the rules which are mined in step 4 to make the final rule base.

Figure 3 shows the method of merging rules for a simple example with 3 levels in partitioning. It shows that the rules in lower level are repeated enough and merge with other rules in higher level to complete the final rule base. For example, the 5th extracted rule is $\mathrm{R}^5 = \left[R_{12} \vdots R_{23} \vdots R_{35}\right]$.
The confidence value of each rule is the arithmetic mean of the confidence values of partial extracted rules. For instance the confidence value of rule 5 is:

$$FC\left(5^{\text{th}} \text{ rule}\right) = \text{mean}\Big(FC\left(\text{rule } R_{12}\right), FC\left(\text{rule} R_{23}\right), FC\left(\text{rule } R_{35}\right)\Big)$$

The final confidence values are calculated after rescaling by the following relation:

$$FC^k = \frac{FC\left(k^{\text{th}} \text{ rule}\right)}{\max\limits_k \left(FC\left(k^{\text{th}} \text{ rule}\right)\right)} \text{ for } k = 1, 2, \ldots, R. \quad (12)$$

5. USING FPM IN PREDICTION PROCEDURE (FPM-PP)

In this study the procedure of the relapse time prediction by fuzzy rules has been divided into four steps called (FPM-PP1, FPM-PP2, FPM-PP3 and FPM-PP4 respectively) as shown in figure 4. In FPM-PP1, the method of gene selection and some needed parameters of FPM algorithm have been recognized and followed by bootstrap procedure for selecting the significant genes in FPM-PP2. Genetic algorithm has been used in an iterative procedure to filter out some of genes from antecedent part of fuzzy rules to increase the accuracy of the prediction in FPM-PP3 and finally in FPM-PP4, the uncertainties of the fuzzy confidence value have been tuned.

5.1 FPM-PP1

Figure 5 shows the flowchart of FPM-PP1. Three methods of gene selection have been considered in the block A to sort the genes. High scored genes (determined by "*ng*" that is added by 5 in each iteration) have been participated in the procedure (in block B).

The "p" levels of partitioning (3, 4 and 5) in FPM have been tested in block C to generate at most 8, 16 and 32 fuzzy rules respectively. Two kernel functions of SVM ("Linear" and "RBF") are applied in this block. Tenfold CV (repeated three times) has been used to estimate the mean of relative errors in prediction with respect to the number of genes and the number of levels. Therefore, for each method of gene selection, the FPM algorithm has been repeated $10 \times 3 \times 17 \times 3 = 1530$times. Two methods of learning procedure (global and combination of global-local method) have been used in block D for calculating the consequent parameters of the fuzzy rules.

Error estimation is calculated in block E. Let $\mathrm{re}_{GS_i}^{p,ng,lm,kf}$ denote the relative error of the models with the method of gene selection GS_i ($i = 1, 2, 3$), ng as the number of selected genes ($ng = 20, 25, \ldots, 100$) in p levels ($p = 3, 4, 5$) of the FPM algorithm where learning method is lm (lm can be global method (denoted by "g")) or combination of global-local method (denoted by "c") and kernel function in SVM is denoted as "*kf*" ("*kf*" can be linear (denoted by "LIN") or RBF (denoted by "RBF")). Then relative error is calculated as follow:

$$\mathrm{re}_{GS_i}^{p,ng,lm,kf} = \frac{1}{3}\sum_{ni=1}^{3}\frac{100}{NS}\sum_{j=1}^{NS}\frac{|RRT\,(j) - Z\,(j)|}{RRT\,(j)} \quad (13)$$

where NS is the number of training samples.

The winner model (selected in block F) is the one (presented in block H) which has minimum $re_{GS_i}^{p,ng,lm,kf}$ denoted by p^*, ng^*, lm^*, kf^* and GS_* where the subset of genes (saved in block G) is named U^*.

The most important role of FPM-PP1 is to determine the best method of gene selection along with many other important parameters such as the number of levels, number of selected genes and so on.

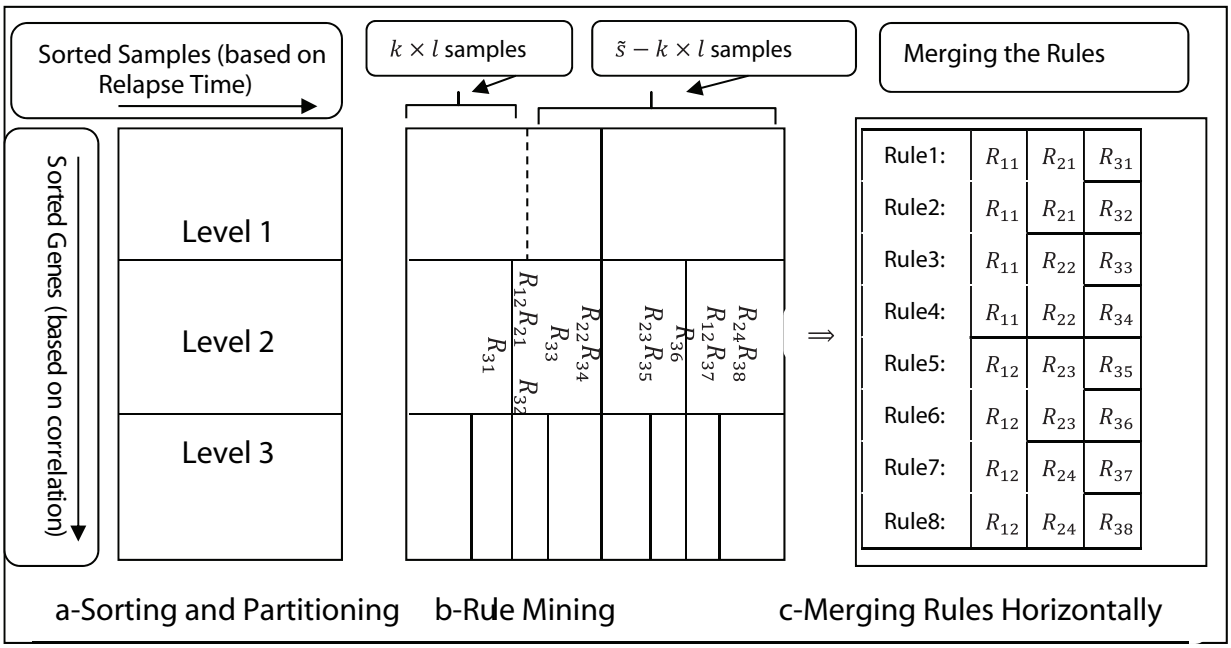

Figure 3 Schematic diagram of the algorithm for a 3-levels example a) row partitioning b) groups and subgroups in the levels and c) merging the rules for making rule base.

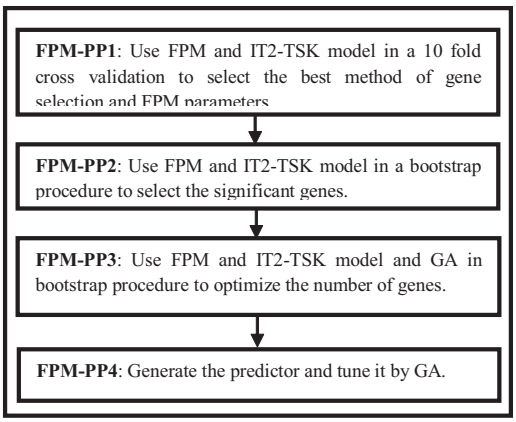

Figure 4 Procedure of the predictor.

5.2 FPM-PP2

To realize the best set of genes based on the output of FPM-PP1, the predictor model has been analyzed again while the number of genes, participated in this step, is added one by one in bootstrap procedure. So the goal of this step is to find the final subset of genes of U^*.

Therefore, selected genes in the U^* have been sorted again based on the best method of gene selection in part 1(GS_*) and IT2-TSK predictor model have been constructed to predict the relapse time of samples to compute the mean of relative error in 60 iterations of bootstrap procedure while the number of genes is added one by one. The subset of genes which has minimum relative error has been selected to be applied to the third part of the procedure.

5.3 FPM-PP3

Definitely, there are some genes with redundant information in the subset of selected genes in FPM-PP3 which may reduce the performance of the model. Genetic algorithm has been used to optimize the number of genes according to the cost function J_{GA} in (14) which is based on mean of Square Correlation Coefficient (SCC) between the real and predicted values.

$$J_{\mathrm{GA}} = \frac{1}{\overline{\mathrm{SCC}} + \varepsilon} \text{ where} \overline{\mathrm{SCC}} = \frac{1}{5}\sum_{i=1}^{5} SCC^i \text{ and}$$
$$\varepsilon = \begin{cases} 0 \overline{SCC} \neq 0 \\ \text{very small value} \overline{SCC} = 0 \end{cases} \tag{14}$$

SCC^i is the square correlation coefficient between predicted and real values of the removed samples in i^{th} loop of 5 fold CV which is internally done in bootstrap procedure. A bit string (0 or 1) is assigned for each gene and the ones with respect to bit "1" are removed to maximize $\overline{SCC}$. The genes which are removed more than a threshold value (τ^* shown in percentage of 50 iterations) have been filtered out. To find the optimumτ^*, an extra 10 fold CV procedure is used in the following steps:

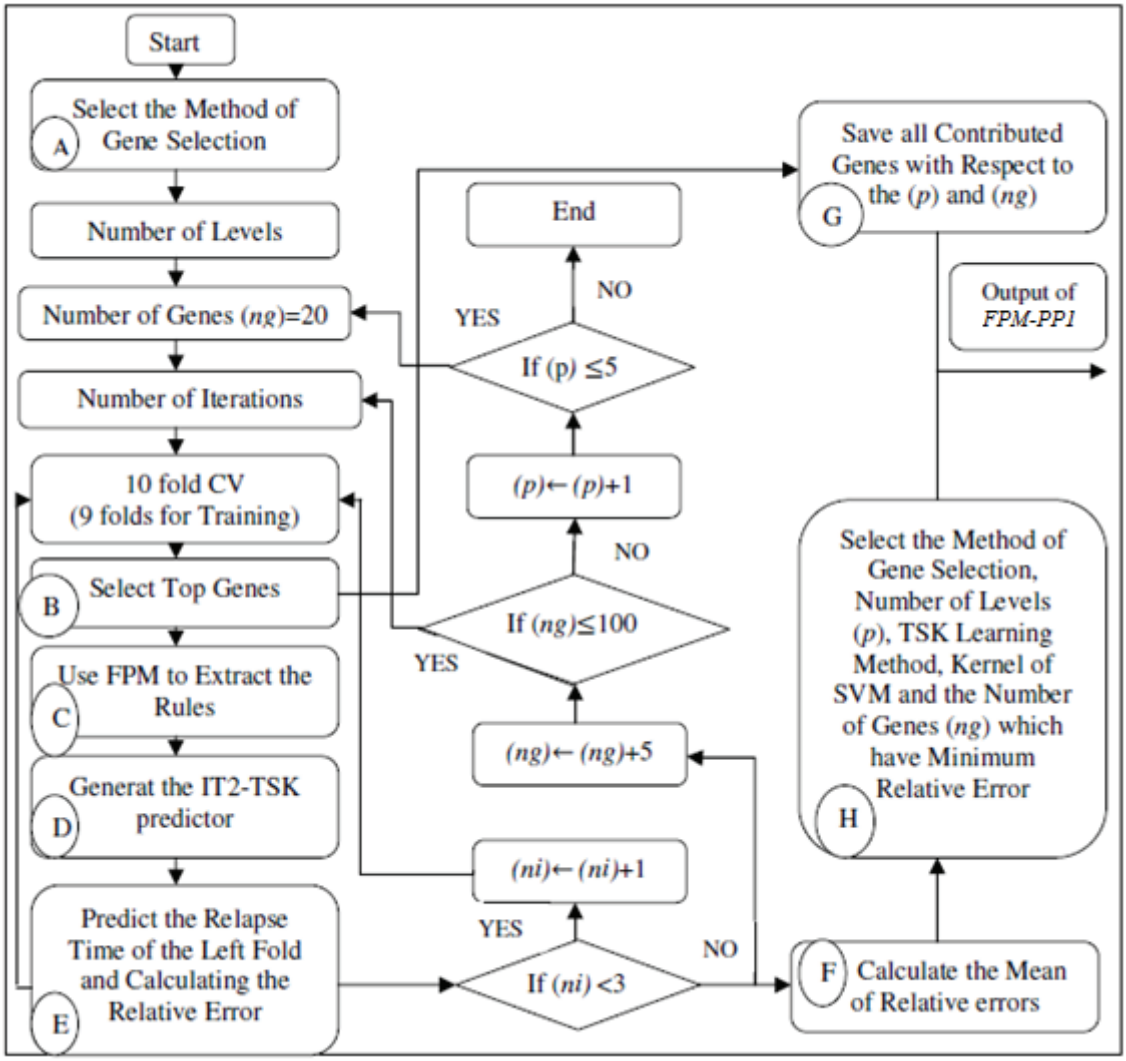

Figure 5 Flowchart of step 1.

1. Set $\tau^* = 30$ (it means that the removed genes in more than or equal 30% of "*nb*" (in figure 6) should be filtered out).

2. Remove out the genes which have been omitted in more than $(\tau^* \times nb)/100$ from the list "L" (in figure 6) and generate a new rule base.

3. Apply 10 fold cross validation procedure on training samples.

1. Use the new rule base and train the model with training samples.

5. Predict the relapse time of the independent samples by the model.

6. CalculateSCC_{cv} (τ^*).

7. $\tau^* \leftarrow \tau^* + 2$and if it is not equal 100 go to step 2, otherwise go to next step.

8. Find τ^* with respect to the maximum SCC_{cv} (τ^*).

5.4 FPM-PP4

To increase the precision of rule base in the predictor, confidence values have been regarded for generated rules of the model by equation (12). It is clear that the confidence value has some dependencies on some conditions such as number of samples or number of levels in FPM. Therefore, there are some uncertainties around the amount of confidence values that should be considered.

In this step the boundary of the uncertainties of fuzzy confidence values have been optimized by another GA procedure to minimize the weighted relative error defined in (15).

$$J_{\text{GA}} = \sum_{j=1}^{s} [\text{RRT}(j) - Z(j)]^2 / \text{RRT}(j). \qquad (15)$$

6. RESULTS AND DISCUSSION

6.1 Results of step 1 (FPM-PP1)

Twelve models (named in Table 1) have been analyzed in FPM-PP1. The results of 10 fold CV of 12 models with respect to the different number of genes and different number of levels

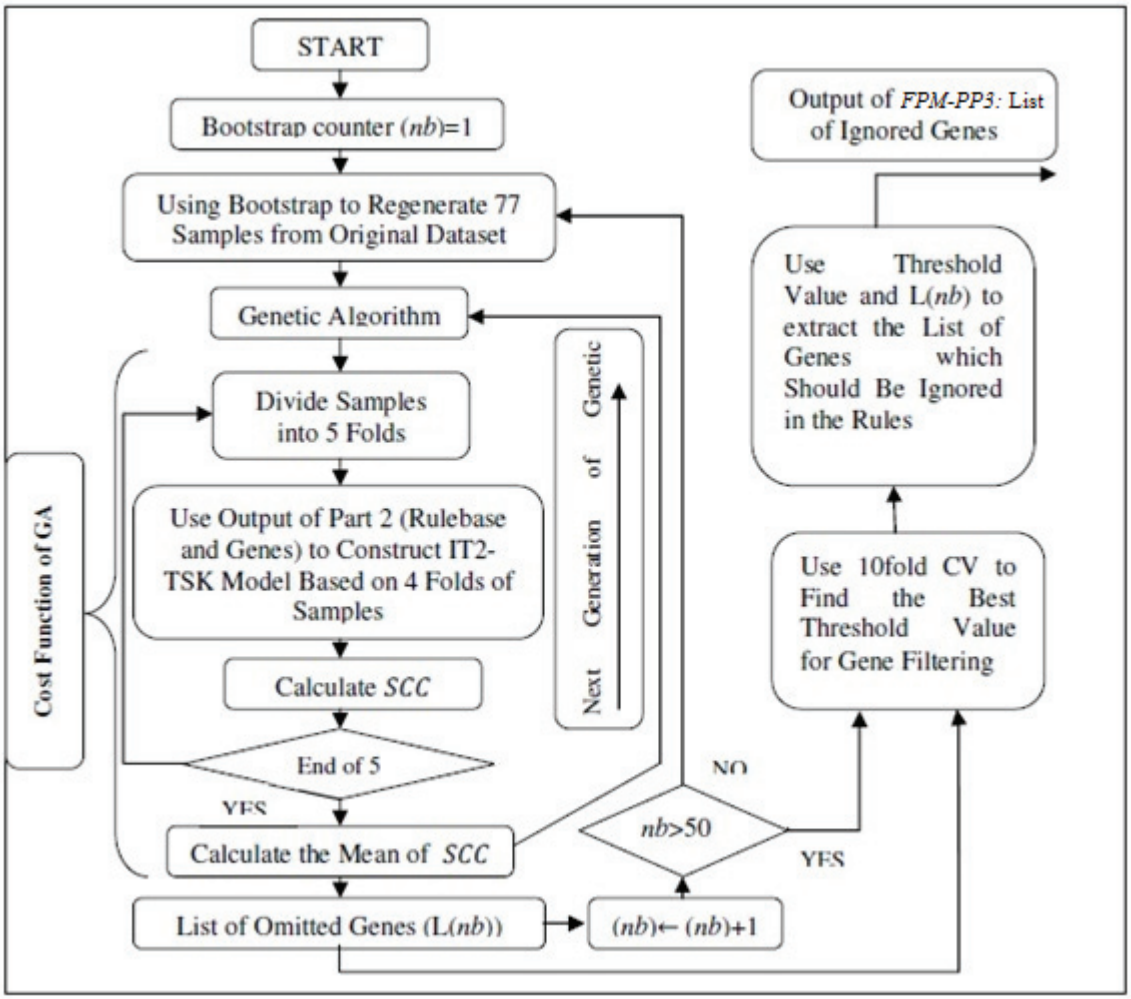

Figure 6 Flowchart of step 3.

(p^*) have been drawn in Figures 7, 8, 9. The figures show that the minimum relative error is related to the PCC_LIN_G with 85 selected genes with 4 levels in FPM ($p^* = 4$).

As it is shown in figure 7(a), the best performance of prediction is achieved by model PCC-LIN-G where the relative error between real and predicted values are about 52.89 with 85 genes. Overall, figures 7, 8 and 9 show that the models with global minimization in coefficient realization of TSK rules and linear kernel in SVM with four levels in partitioning, have a better performance in prediction. The results of this step convince us to optimize the performance of the model based on PCC-LIN-G.

6.2 Results of step 2 (FPM-PP2)

In this part, all 194 genes contributed to 30 iterations of cross validation (collected in the subset U^*) have been used to predict the 77 original samples in a B632 procedure based on PCC-LIN-G while the genes were added one by one in this step. As depicted in Figure 10, the best performance has been achieved with 87 high correlated genes (55.37).

The generated rule base with 87 selected genes includes 16 rules which separate the 77 samples into 16 subgroups as shown in Figure 11 and table 2. In Figure 11, the color map of the gene expression values of 87 sorted genes in 77 sorted samples have been presented so that the maximum gene expression value (4.076) and minimum gene expression value (-6.64) have been normalized between 3 (red color) and -3 (blue color) respectively. The black lines show the boundaries between the subgroups in 4 levels of feature partitioning in the FPM algorithm.

Table 2 shows that FPM algorithm can gather the samples into the three major groups: under 2.5 years (very high risk sample), between 2.6 and 4.89 years (high risk samples) and between 6.25 and 10.23 (low risk samples). For example to analyze the differences between the behavior of genes in group 1 and group 5 (in Table 2), the mean of gene expression values of the samples in two groups 1 and 5 have been plotted in figure 12.

Statistical T-test (Two Samples) is used to analyze the differences between the mean of gene expression values. Let the mean of gene expression values in two groups 1 and 5 be denoted by $\overline{m}_i^1$ and $\overline{m}_i^5$ ($i = 1$ to 87) respectively, so the null and alternative hypothesis could be defined as:

$$H_0 : \overline{m}_i^1 = \overline{m}_i^5$$
$$H_1 : \overline{m}_i^1 \neq \overline{m}_i^5$$

The results of T-test and their p-values have been presented

Table 1 Different categories used in this paper.

Model name	Gene sorting method	Type of Kernel function in SVM (applied in the rule mining)	β_{Global}	β_{Local}
PCC_LIN_G	PCC	linear	1	0
PCC_LIN_C	PCC	linear	0.5	0.5
PCC_RBF_G	PCC	RBF	1	0
PCC_RBF_C	PCC	RBF	0.5	0.5
SNR_LIN_G	SNR	linear	1	0
SNR_LIN_C	SNR	linear	0.5	0.5
SNR_RBF_G	SNR	RBF	1	0
SNR_RBF_C	SNR	RBF	0.5	0.5
SVM_LIN_G	SVM-RFE	linear	1	0
SVM_LIN_C	SVM-RFE	linear	0.5	0.5
SVM_RBF_G	SVN-RFE	RBF	1	0
SVM_RBF_C	SVM-RFE	RBF	0.5	0.5

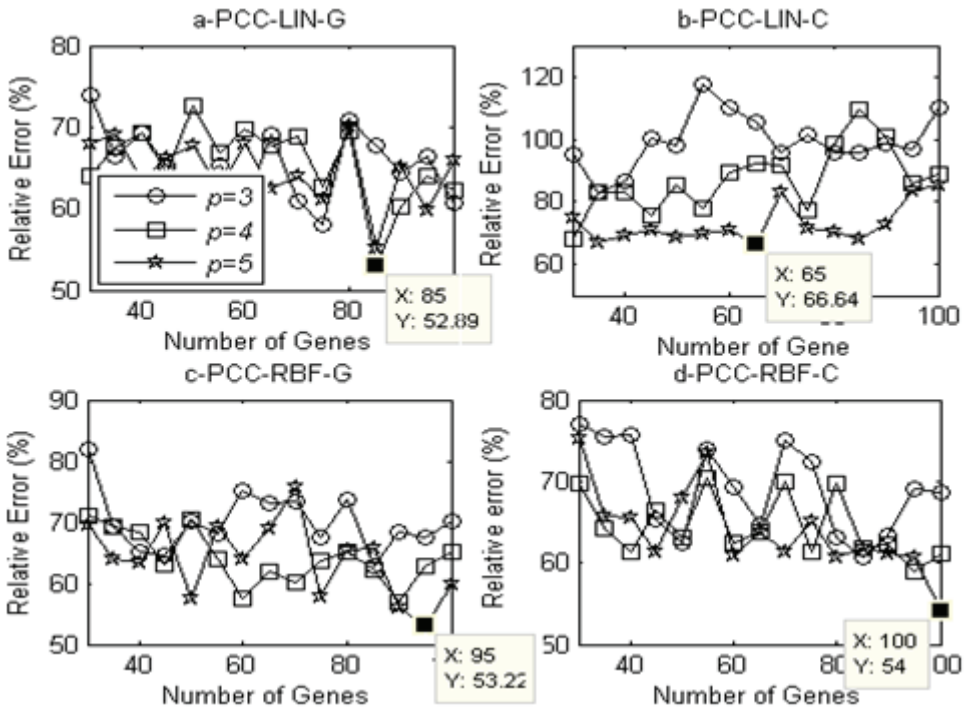

Figure 7 Results of four models of rule mining and rule learning with PCC-Continuous Scale.

in Table 3 and show that the null hypothesis was rejected for 18 of 87 genes with 95% significant level. The mean of gene expression values of four genes (bolded ones in Table 3) have similar sign which shows similarity in regulation (16^{th} gene is up-regulated and 18^{th}, 60^{th} and 61^{th} are down-regulated). In such case, linguistic variables (defined in fuzzy rules) have more ability to separate the groups comparing with binary up-down regulation.

6.3 Results of step 3 (FPM-PP3)

To improve the performance of the model and reduce the number of genes, genetic algorithm has been used in FPM-PP3 as explained in step 3. The number of times that 87 genes are removed from the rule base are depicted in Figure 13. Maximum and minimum numbers belong to the genes "X05610" and "NM_000320" with 82% and 14% respectively. NM_000320 with gene symbol QDPR (Quinoid Dihydropteridine Reductase) has been known earlier as a good prognostic marker in breast cancer [31].

Another 10 fold cross validation procedure has been applied to specify the best threshold level (τ^*).Figure 14(a) shows that the maximum *SCC* has been achieved when the threshold value τ^* was equal to 52% (26 of 50 times). By this amount, the number of significant genes in the model is reduced to 46 (Figure 14(b)).

The new rule base (include 46 genes) has been performed and the new model has been analyzed with independent samples in two following approaches (results are shown in Table 4).

1. Predicting the relapse time of two independent datasets with different trained samples (generated in a 10 fold

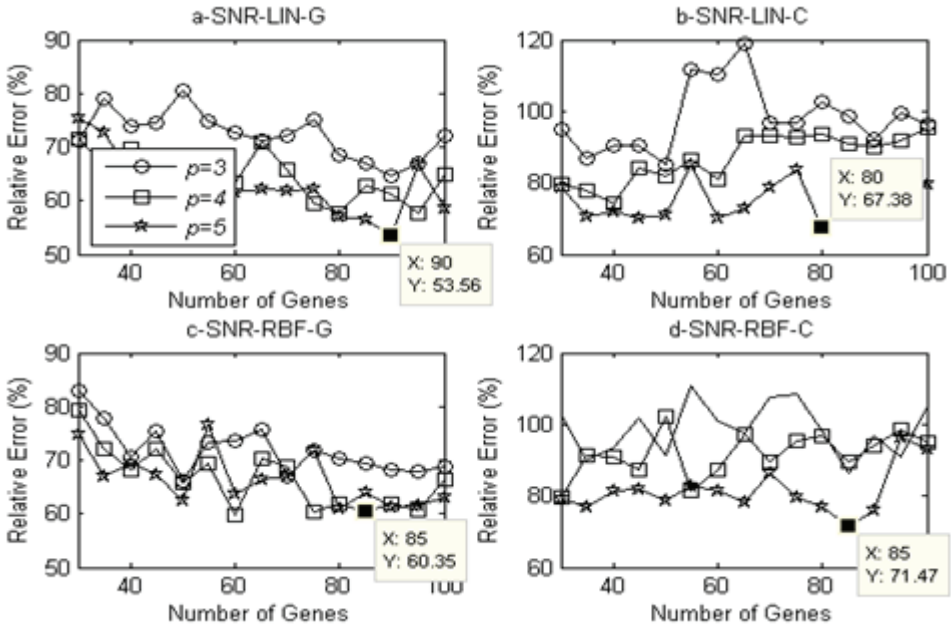

Figure 8 Results of four models of rule mining and rule learning with SNR.

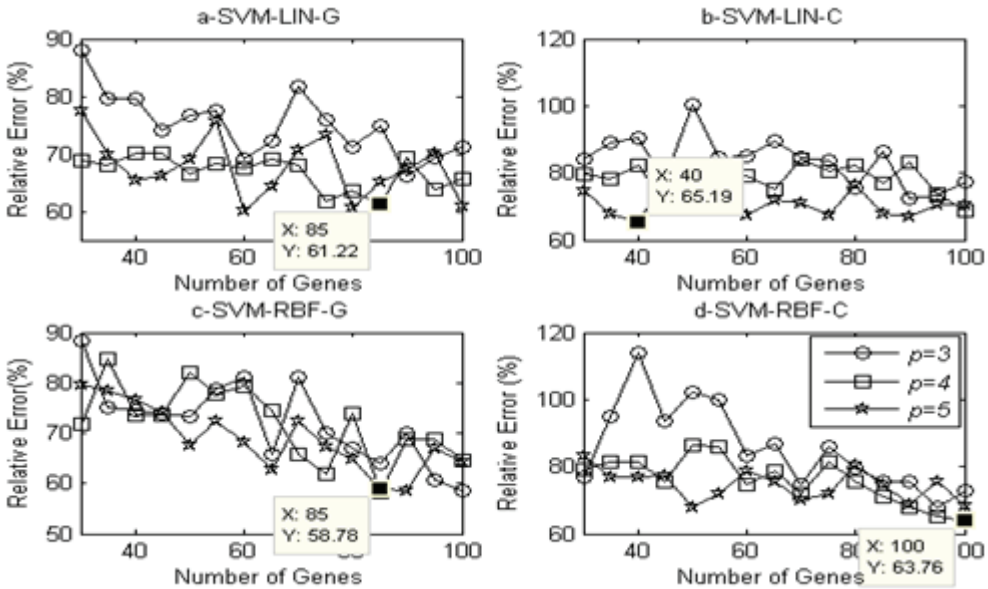

Figure 9 Results of four models of rule mining and rule learning with SVM-RFE.

CV and repeated 5 times) to test the reliability of the model. It is noticeable that different training samples generate different values for the consequent parameters of fuzzy rules. Mean of *SCC's* and relative errors have been denoted respectively by $\overline{\mathrm{SCC}}_{19}$ and $\overline{\mathrm{RE}}_{19}$ for 19 samples, $\overline{\mathrm{SCC}}_{36}$ and $\overline{\mathrm{RE}}_{36}$ for 36 samples.

2. Predicting the relapse time of two independent samples when the model has been trained with 77 original samples. *SCC's* and relative errors of two datasets have been denoted by $SCC_{19}$and RE_{19} for 19 samples, $SCC_{36}$and $RE_{36}$for 36 samples respectively.

The results of the proposed algorithm and the presented results in [13] have been compared in the Table 5.

Although the number of selected genes in our model is more than proposed model in [13], the results indicate that the proposed model based on FPM algorithm has superiority in predicting the relapse time in comparison with previous work. The results of prediction with linear regression line of two independent datasets are shown in Figure 15.

To compare the IT2-TSK model with a similar T1-TSK model, all uncertainties (u_t^k) have been considered equal to zero and two independent samples have been predicted in 10 fold CV procedure which was repeated 5 times. Let denote the mean of relative errors of both 19 and 36 datasets by $\overline{m}_{T2}$ and $\overline{m}_{T1}$ for IT2-TSK and T1-TSK respectively, the following

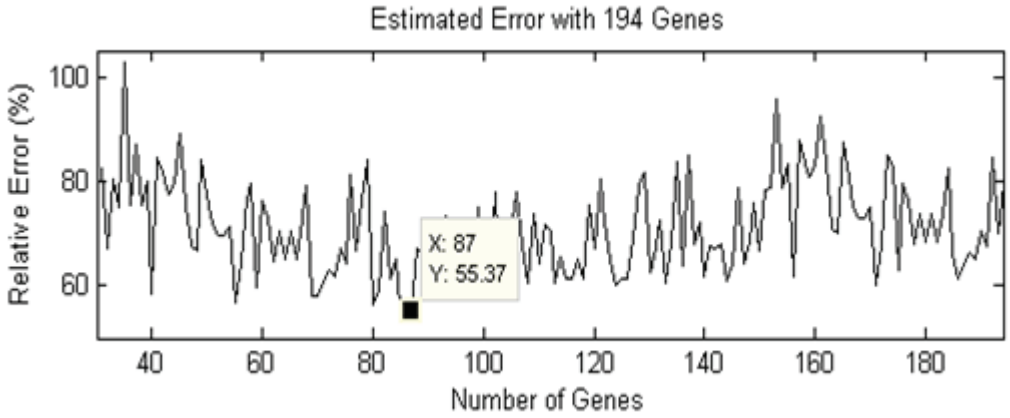

Figure 10 Result of Prediction with model PCC-LIN-G.

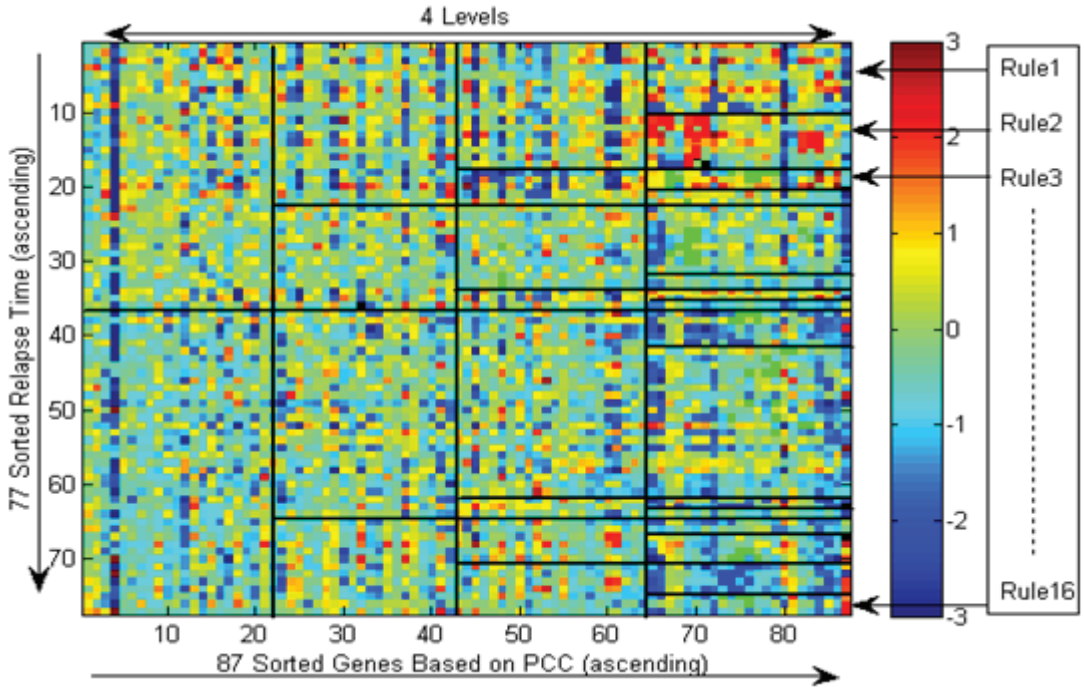

Figure 11 Color Map of 87 Genes in 77 Sample and 16 Generated Subgroups

hypothesis has been tested by T-test (Two Samples).

$$H_0 : \overline{m}_{T2} \geq \overline{m}_{T1}$$
$$H_1 : \overline{m}_{T2} < \overline{m}_{T1}$$

The null hypothesis was rejected by 90% CI that shows the superiority of IT2-TSK to T1-TSK, although $\overline{m}_{T2}$ and $\overline{m}_{T1}$ were equal to 64.65 and 69.09 respectively.

6.4 List of selected genes

The list of 46 selected genes is in table 6. Seven genes Contig63102_RC, Contig40831_RC, NM_004508, NM_006681, NM_000320, Contig33814_RC and Contig42421_RC are common among two subsets of selected genes in this research and [13]; while 17 genes of the subset have been reported in [32] as high impact genes in breast cancer outcome and in the dataset of 70 genes reported by van't Veer. Moreover, 5 of them (third column in table 6) have been considered as good prognosis markers in the supplementary data of [31]. Although 27 of 46 genes have been reported in previous researches as high impact markers in outcome of breast cancer, more experiments with more samples are necessary to confirm the effect of other 19 genes in breast cancer outcome.

6.5 Results of step 4 (FPM-PP4)

Uncertainties of the confidence values of the rules (u^k_{virtual}, $k = 1, 2, \ldots, 16$) have been calculated by using GA in a 10 fold CV. Equation (15) has been considered as the cost function to improve two following goals in the model:

1. Decreasing the error of prediction:

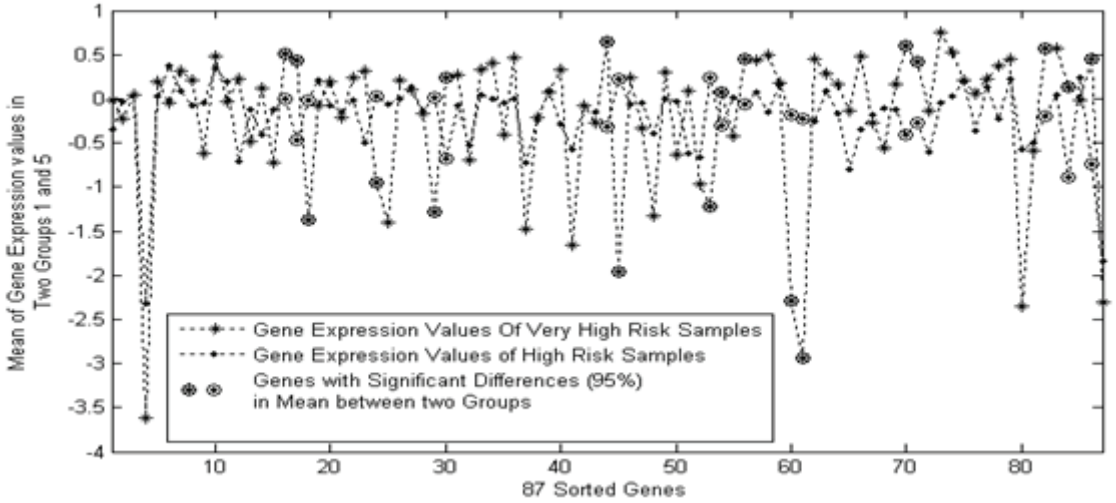

Figure 12 Mean of Gene Expression values in two subgroups 1 and 5.

Table 2 16 Generated Subgroups by FPM Algorithm.

Groups	**The number of samples in the groups**	**relapse time intervals**
Group 1	10	0.27-1.62
Group 2	8	1.85-2.15
Group 3	2	2.21-2.3
Group 4	2	2.31-2.68
Group 5	10	2.68-4.89
Group 6	2	4.9-5.1
Group 7	1	5.18-5.18
Group 8	1	5.19-5.19
Group 9	6	5.2-5.89
Group 10	20	6.25-10.23
Group 11	1	10.29-10.29
Group 12	1	10.38-10.38
Group 13	2	10.54-10.66
Group 14	4	11.02-11.46
Group 15	4	11.84-12.53
Group 16	3	12.73-13.42

Since the square correlation coefficient has been considered in the cost function of GA in the part 3, in this part sum of weighted square errors of the prediction has been applied in cost function (15) to decrease the error of the model.

1. Decreasing the false negative errors:

In equation (15) the square of the errors have been weighted by the parameter $1/RRT\,(j)$ where $RRT\,(j)$ is the real relapse time of j^{th} sample. By this weighting method, the penalty on the samples with smaller real relapse time is more than the samples with longer one.

Table 7 shows the confidence values of the rules and their uncertainties.

To analyze the impact of cost function on the false negative error, the sum of square of the errors in two groups of samples (within 5 and after 5 relapse time) of 19 and 36 datasets have been calculated separately with and without applying the uncertainty of the rules.

Suppose that $\overline{\text{MSE}} = \frac{1}{ns}\sum_j (\text{RRT}\,(j) - Z\,(j))^2$ be the mean of square of errors ($\overline{\text{MSE}}$) where "*ns*" is the number of samples, then $\overline{\text{MSE}}_{\text{un}}^{<5}$, $\overline{\text{MSE}}_{un}^{>5}$, $\overline{\text{MSE}}_{no}^{<5}$ and $\overline{\text{MSE}}_{no}^{>5}$ are denoted as MSEs of the two groups of samples which have been predicted with and without uncertainty in the model. ("<5" and ">5" are denoted for the samples with less than and greater than 5 years relapse time respectively). Let define $\xi^{<5}$ and $\xi^{>5}$ as the percentage of growth in MSE in the created model at part 3 with respect to the created model at part 4 with following equations:

$$\xi^{<5} = \frac{\overline{\text{MSE}}_{no}^{<5} - \overline{\text{MSE}}_{un}^{<5}}{\overline{\text{MSE}}_{no}^{<5}} \times 100 \qquad (16)$$

$$\xi^{>5} = \frac{\overline{\text{MSE}}_{no}^{>5} - \overline{\text{MSE}}_{un}^{>5}}{\overline{\text{MSE}}_{no}^{>5}} \times 100 \qquad (17)$$

To analyze the hypothesis which states: "is $\xi^{<5}$ less than $\xi^{>5}$ or no?" the samples of both 19 and 36 independent datasets have been predicted and their MSEs ($\overline{\text{MSE}}_{un}^{<5}$, $\overline{\text{MSSE}}_{un}^{>5}$, $\overline{MSE}_{no}^{<5}$, $\overline{MSE}_{no}^{>5}$) were calculated while the training samples were selected based on 10 fold CV which has

Table 3 Differences in Mean of Gene Expression Values in Subgroups 1 and 5.

i gene in 87 sorted genes	Name of Genes with significant differences in two groups	m_i	m_i	p-value
16	**NM_006027**	**0.51**	**0.005**	**0.03**
17	Contig37024_RC	0.43	-0.46	0.04
18	**NM_002051**	**-1.36**	**-0.01**	**0.04**
24	NM_003022	-0.94	0.035	0.02
29	NM_005375	-1.28	0.017	0.04
30	AF113132	0.23	-0.67	0.01
44	D82345	0.63	-0.31	0.02
45	Contig56390_RC	-1.96	0.22	0.009
53	U45975	-1.21	0.24	0.001
54	NM_004504	0.07	-0.30	0.01
56	D55716	0.44	-0.05	0.01
60	**Contig43983_RC**	**-2.29**	**-0.17**	**0.008**
61	**Contig47405_RC**	**-2.93**	**-0.22**	**0.006**
70	Contig28552_RC	0.59	-0.40	0.04
71	NM_004336	0.42	-0.27	0.03
82	AL137718	0.57	-0.19	0.04
84	NM_001333	0.14	-0.89	0.03
86	NM_006681	0.44	-0.73	0.03

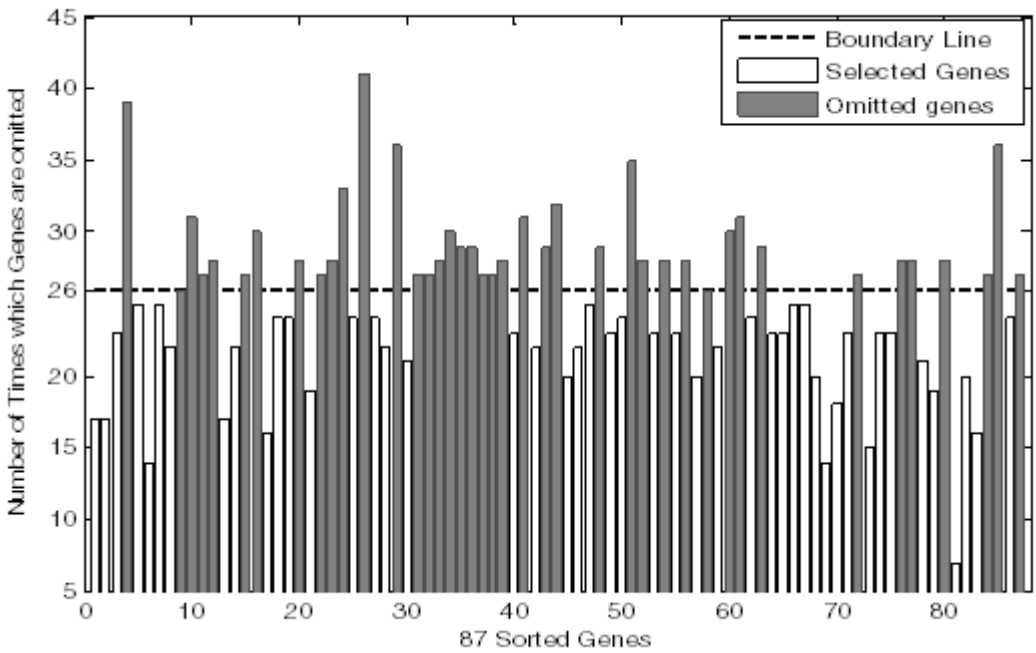

Figure 13 Results of 50 iterated GA in Part 3.

Table 4 Results of Prediction of Independent Samples in Part3.

Dataset	$\overline{SCC}$	SCC	$\overline{RE}$(%)	RE(%)
19 samples	$\overline{SCC}_{19}$=0.552	SCC_{19}=0.698	$\overline{RE}_{19}$=58.52	RE_{19}=57.11
36 samples	$\overline{SCC}_{36}$=0.431	SCC_{36}=0.64	$\overline{RE}_{36}$=73.10	RE_{36}=63.47

been repeated 5 times. Therefore, there are 50 values for $\xi^{<5}$ and $\xi^{>5}$ which is analyzed by T-test with the following one-tail hypothesis.

$$H_0 : \overline{\xi^{<5}} \geq \overline{\xi^{>5}}$$
$$H_1 : \overline{\xi^{<5}} < \overline{\xi^{>5}}$$

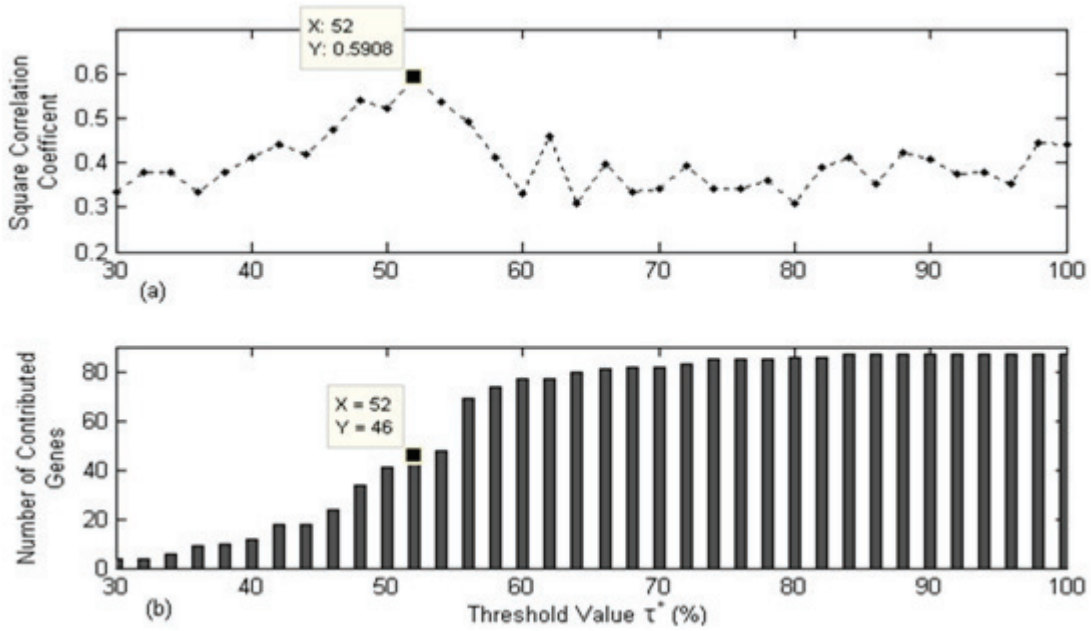

Figure 14 Results of 10 Fold CV in Part 3 to find the Best Subset of Genes (a) maximum SCC (b) Number of genes in the rule base.

Table 5 Comparison of the Results between Proposed Algorithm and Previous Research.

Samples	Prediction by IT2-TSK mode l	Prediction by regression in [13]
Estimated $\boldsymbol{SCC}$ in the training dataset	0.625 with 77 samples in the original dataset	0.57 and 0.58 with respect to two training dataset including 97 and 74 samples
$\boldsymbol{SCC}$ of independent datasets	0.698 (SCC_{19}) and 0.640 (SCC_{36}) with respect to the two independent datasets including 19 and 36 samples	0.62 for a dataset including 23 samples (but some samples have been used in the gene selection procedure)
Number of genes	46	41

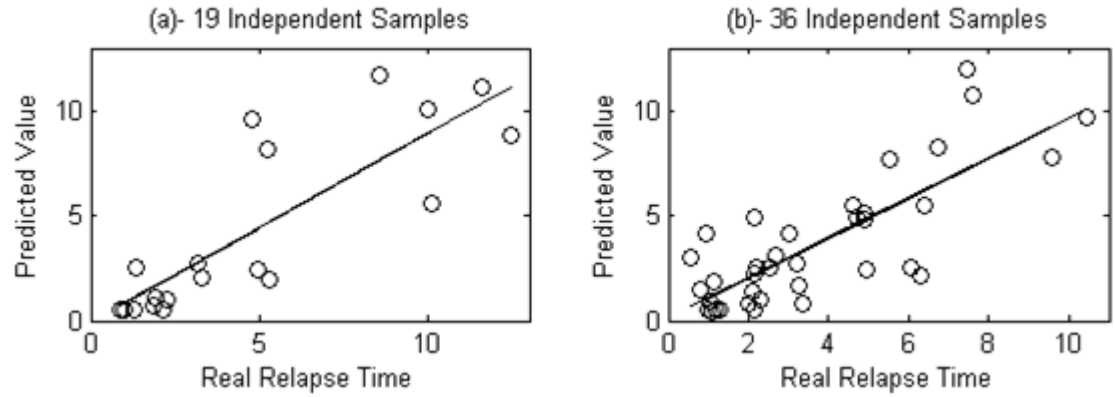

Figure 15 Predicted values of Independent Samples in Part 3 (a) 19 samples of van't Veer Dataset (b) 36 Samples of Vijver Dataset.

Where $\overline{\xi^{<5}}$ and $\overline{\xi^{>5}}$ are the mean of $\xi^{<5}$ and $\xi^{>5}$ respectively. The result of T-test confirmed that the null hypothesis was rejected with 95% CI. In this experiment, $\overline{xi^{<5}}$ and $\overline{\xi^{>5}}$ were 1.37 (%) and 3.52(%) respectively which shows that the false negative error is reduced in the final model, although the consideration of the rule uncertainties had no impact on overall relative errors which is analyzed in the previous hypothesis.

Finally, the relapse time of the samples in both independent datasets have been predicted by two following approaches:

1. By training the model in 10 fold CV procedure which has been repeated 10 times. In this approach, the mean of square correlation coefficient and relative errors in 10 repetitions have been denoted with $\overline{SCC}_{19}$, $\overline{SCC}_{36}$, $\overline{RE}_{19}$ and $\overline{RE}_{36}$ and then presented in Table 8. Figure 16 shows the predicted values with 95% CI.

2. By training the model with 77 original samples of van't Veer dataset. SCC_{19}, SCC_{36}, RE_{19} and RE_{36} are shown in Table 8; predicted values and regression line are pre-

Table 6 Description of the 46 Selected Genes.

	Name of Genes	Reported in other research	Gene ontology
g_1	Contig63102_RC	[32], [13] ,[1]	
g_2	NM_016337	[31]	SH3 domain ,binding, actin binding, profilin binding, protein bindin
g_3	Contig40831_RC	[32],[1]	
g_4	NM_003875	[32], [52],[1]	ATP Binding-GMP synthase Activity-Enzyme Binding- Ligase Activity Nucleotide Binding
g_5	Contig2399_RC	[32],[1]	L-Ascorbic Acid Binding- Iron Ion Binding-Metal Ion Binding
g_6	NM_020188	[32],[1]	
g_7	NM_006304		Peptidase activity, protein binding
g_8	NM_018166		Protein binding
g_9	Contig44799_RC		
g_{10}	Contig37024_RC		
g_{11}	NM_002051		DNA binding ,metal ion binding sequence-specific ,DNA binding transcription factor activity, transcription factor binding zinc ion binding
g_{12}	NM_004508	[13]	hydrolase activity, isomerase activity, isopentenyl-diphosphate delta-isomerase activity, metal ion bindin
g_{13}	Contig35076_RC		
g_{14}	AB020689	[33]	GTPase activator activity , Rab GTPase activator activity , calcium ion binding
g_{15}	NM_000291	[31]	ATP binding , nucleotide binding phosphoglycerate kinase activity , transferase activity
g_{16}	Contig50802_RC		
g_{17}	AF113132		O-phospho-L-serine:2-oxoglutarate aminotransferase activity pyridoxal phosphate binding transferase activity
g_{18}	U96131		ATP binding , identical protein binding nucleoside-triphosphatase activity , nucleotide binding protein binding transcription cofactor activity
g_{19}	NM_020386	[32],[1]	Cytoplasm, nuclear envelope lumen
g_{20}	Contig56390_RC		
g_{21}	Contig38288_RC	[32],[1]	
g_{22}	Contig43747_RC		
g_{23}	NM_015239		Binding- Metal Ion Binding- Metallocarboxypeptidase Activity- Peptidase Activity- Zinc Ion Binding
g_{34}	NM_015417		Molecular function
g_{25}	U45975		SH3 domain binding, hydrolase activity, inositol or phosphatidylinositol phosphatase activity inositol-polyphosphate 5-phosphatase activity
g_{26}	Contig42421_RC	[13]	
g_{27}	NM_005915	[32],[1]	ATP binding ,PubMed DNA binding , DNA helicase activity , identical protein binding nucleotide binding protein binding single-stranded DNA binding
g_{28}	Contig63649_RC	[32],[1]	
g_{29}	Contig46218_RC	[32],[1]	
g_{30}	NM_013296		GTPase activator activity , identical protein binding protein binding
g_{31}	NM_018265	[31]	
g_{32}	Contig55725_RC	[32],[1]	
g_{33}	U82987	[32],[1]	Protein binding
g_{34}	NM_006763		Protein binding
g_{35}	NM_018685		Actin binding
g_{36}	Contig28552_RC	[32],[1]	
g_{37}	NM_004336	[31]	ATP Binding- Nucleotide Binding
g_{38}	Contig25991	[31],[1]	Rho GTPase Binding- Protein Binding
g_{39}	NM_017779		GTPase Actiator Activity
g_{40}	AL050151		
g_{41}	NM_001673	[34]	ATP binding , asparagine synthase (glutamine-hydrolyzing) activity, asparagine synthase (glutamine-hydrolyzing) activity cofactor binding ligase activity , nucleotide binding protein homodimerization activity
g_{42}	Contig20217_RC	[32],[1]	
g_{43}	NM_000320	[31], [13]	6,7-dihydropteridine reductase activity binding, electron carrier
g_{44}	AL137718	[32],[1]	Actin binding
g_{45}	Contig33814_RC	[13]	
g_{46}	NM_006681	[32], [13],[1]	Neuromedin U Receptor Binding- Receptor Binding

sented in Figure 17.

7. CONCLUSION

In this study, the prediction of relapse time of breast cancer in a continuous scale was considered and it shows that a fuzzy model based on T2-TSK, which is able to handle the uncertainty in linguistic variables, have superiority in prediction with respect to a regression model presented in previous research. In the proposed algorithm, GA has been used to optimize the number of genes which are related to the relapse time and most of them have also confirmed earlier in previous studies.

Table 7 Confidence Values and their Uncertainties of 16 Rules.

	Confidence value	Uncertainty		Confidence value	Uncertainty
Rule 1	1	0	Rule 9	0.41	±0.20
Rule 2	0.8	±0.14	Rule 10	0.39	±0.17
Rule 3	0.41	±0.11	Rule 11	0.31	±0.14
Rule 4	0.71	±0.24	Rule 12	0.42	±0.13
Rule 5	0.78	±0.12	Rule 13	0.62	±0.13
Rule 6	0.71	±0.27	Rule 14	0.64	±0.22
Rule 7	0.21	±0.11	Rule 15	0.57	±0.05
Rule 8	0.33	±0.05	Rule 16	0.51	±0.25

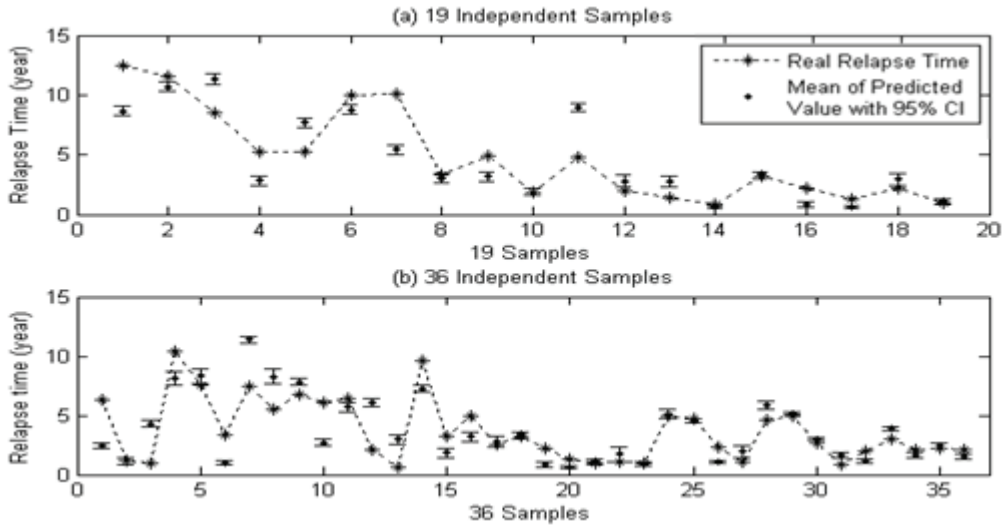

Figure 16 Mean of Predicted Relapse Time and 95%CI in Part 4 (a) 19 Independent Samples (b) 36 Independent Samples.

Table 8 Results of Prediction of Independent Samples by Final Model.

19 independent samples		36 independent samples	
$\overline{SCC}_{19}$	$\overline{RE}_{19}$(%)	$\overline{SCC}_{36}$	$\overline{RE}_{36}$(%)
0.62	54.96	0.49	71.79
SCC_{19}	RE_{19}(%)	SCC_{36}	RE_{36}(%)
0.71	41.40	0.64	59.95

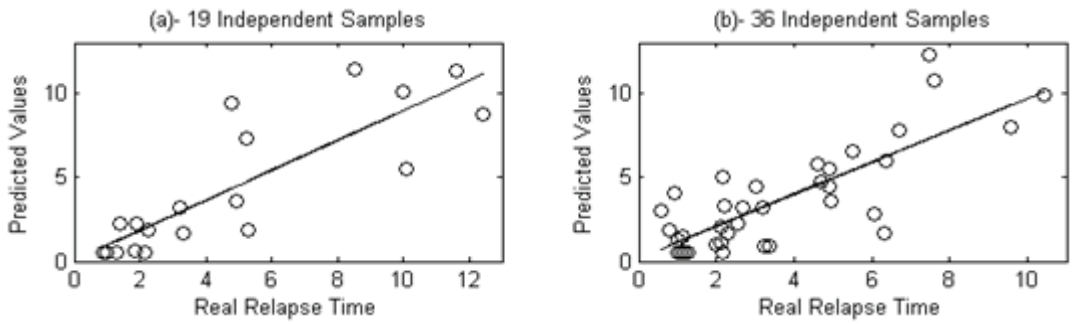

Figure 17 Predicted values of Independent Samples in Part 4 (a) 19 samples of van't Veer Dataset (b) 36 Samples of Vijver Dataset.

Overall, survival analysis by gene expression profile are including many challenges that we tried to consider them in our study. The challenges such as existence of a lot of genes in the profile with uncertainty on their gene expression, redundancy between them and or finding how the genes are related to each other and how we can state these relations by linguistic variables. In spite of high computational cost, in this study different cost function such as relative error, weighted relative error and square correlation coefficient have been minimized by genetic algorithm and then statistically analyzed. In addition, to decrease the over fitting problem and to increase the reliability of the model, cross validation and bootstrap procedure have been performed in the steps of proposed procedure. Finally, comparing with previous studies, the list of selected

genes have been introduced in this article. In future work, we will work on how we can generate fuzzy rule in simpler way and or how the computational cost could be reduced.

Acknowledgement

This work has been financially supported by Islamic Azad University, Najafabad Branch under grant No. 1349. Any opinion, finding and conclusion expressed in this material are those of the authors and do not necessarily reflect those of the Islamic Azad University, Najafabad Branch.

REFERENCES

1. van't Veer, L, Dai, H,etal (2002). Gene expression profiling predicts clinical outcome of breast cancer.*Nature,* 415: 530–536
2. Alba, E, Garcia-Nieto, J, et al (2007). *Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms*. In: Evolutionary Computation. CEC 2007.IEEE Congress, Date:25-28 Sept.pp: 284–290.
3. Guali, Ji, Zijiang, Y, Wenjie, Y, (2011). PLS-based gene selection and identification of tumor-specific genes. *System, Man, and Cybernetics-Part C: Application and Review*, 41(6).
4. Yu, L, Han, Y, E.Berens, M, (2012). Stable gene selection for microarray data via simple weighting. *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, 9(1).
5. Sipola, T, Juvonen, A, and Lehtonen, J. (2012). Dimensionality reduction framework for detecting anomalies from network logs. *Engineering Intelligent Systems*, 20 (1/2), 87–97.
6. Gruvberger, S, Ringner, M, et al. (2001). Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns.*Cancer Research* 61(15) 5979–5984.
7. Antoniou, P, Michalakopoulos, S, Ionnides, M, A Papageorgiu, E, and C?Patsalis, P (2014). Identifying differentially methylated regions by efficient bit-masking of DNA microarrays to use as markers for non-invasive prenatal diagnosis. *Engineering Intelligent Systems*, 22(2).
8. Jiang, D, Zhao, N, (2006). A clinical prognostic prediction of lymph node-negative breast cancer by gene expression profiles.*Cancer Research Clinical Oncology*, 132(9) 579–587.
9. Alexe, G,Alexe, S. et al, (2005). Breast cancer prognosis by combinatorial analysis of gene expression data.*Breast Cancer Research*. 8(4).
10. Shen, R, Ghosh, D, et al. (2006). Eigengene-based linear discriminant model for tumor classification using gene expression microarray data.*Bioinformatics*, 22(21) 2635–2642.
11. Mahmoodian, H, Hamiruce Marhaban, M, Abdulrahim, R, Rosli, R, and Saripan, I, (2011). Using fuzzy association rule mining in cancer classification. *Australasian Physical and Engineering Sciences in Medicine*, 34(1)41–54.
12. Mahmoodian, H, (2012). Predicting the continuous values of breast cancer relapse time by type-2 fuzzy logic system. *Australasian physical & engineering sciences in medicine,* 35:193–204.
13. Shih-Hau, C., Chien-Chi, C. Thy-Hou, L. (2008). Using support vector regression to model the correlation between the clinical metastases time and gene expression profile for breast cancer. *Artificial Intelligence in Medicine*,44 221–231.
14. Gruvberger, S, Eden, P, et al, (2004). Predicting continuous values of prognostic markers in breast cancer from microarray gene expression profiles.*Molecular Cancer Therapeutics,* 3(2) 161–168.
15. Zade, L, (1976). Fuzzy-algorithm approach to the definition of complex or imprecise concepts. *International journal. Man Mach. Stud,* 8 249–291.
16. Quinlan, R, (1993). C4.5: programs for machine learning. San fransisco; CA, Morgan kaufman.
17. Pal, K, Mitra S, (1999). Neuro-fuzzy pattern recognition: methods in soft computing. Wiely Series on intelligent systems; New York, Wiely- interscience.
18. Kouk, C, Fu, A, et al. (1998). Mining fuzzy association rules in data base. *ACM SIGMOD*, 27(1) 41–46.
19. Chen, S,Wang J, et al. (2008). *Extraction of fuzzy rules by using support vector machines*, Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 3 438–442.
20. Farquad, M, Ravi, V, et al.,(2008). Rule Extraction using Support Vector Machine Based Hybrid Classifier. *TENCONIEEE region 10 conference*, Nov 1–6.
21. Van de vijver, Mj, Yudong, D. et al. (2002). A Gene-Epression Signature as a Predictor of Survival in Breast Cancer.*The New England Journal of Medicine,*347(25).
22. Guyon, I, Weston, J. et al. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46389–422.
23. Mendel, J, (2007). Type-2 fuzzy sets and systems: an overview.*IEEE computational intelligence magazine*, 2(2) 20–29.
24. Karnik, N, Mendel, J, et al. (1999). Type-2 fuzzy logic systems. *IEEE Transaction on Fuzzy Systems*, 7(6).
25. Castillo, O, Melin, P, (2008). *Type-2 Fuzzy Logic: Theory and Applications.* Studies in Fuzziness and Soft Computing, 223. Springer.
26. Karnik, N, Mendel, J, et al. (2001). Centroid of a type-2 fuzzy set. *Information sciences*, 132(1) 195–220.
27. Mendel, J, (2001). *Uncertain rule-based fuzzy logic systems: introduction and new directions*, Printice–Hall, upper-saddle river, NJ.
28. Ishibuchi, H, Yamamoto, T, (2005). Rule weight specification in fuzzy rule-based classification systems.*IEEE Transaction on Fuzzy Systems*, 13(4) 428–435.
29. Vanícek, J, Vrana I, et al. (2009). Fuzzy aggregation and averaging for group decision making:A generalization and survey, *Knowledge-Based Systems* 22(1) 79–84.
30. Yen, J, Wang, L, et al. (1998). Impoving the interpretability of TSK fuzzy models by combining global and local learning. *IEEE Transactionon Fuzzy Systems,* 6(4) 530–537.
31. Fan, M, Yan, P, Hartman-Frey, C, Chen, L, Paik, H, Oyer, S, Salisbury, J, Cheng, A, Li, L, Abbosh, P, Huang, Tand Nephew, K. (2006). Diverse gene expression and DNA methylation profiles correlate with differential adaptation of breast cancer cells to the antiestrogens tamoxifen and fulvestrant. *Cancer Research.*66(24) 11954–11966.
32. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=hserta &part=A257712, "Impact of gene expression profiling tests on breast cancer outcomes
33. Mahata, P, (2010). Exploratory consensus of hierarchical clustering for melanoma and breast cancer.*IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 7(1) 138–152.
34. Stoddard, F, Brooks, A, Eskin, B, Johannes, G, (2008). Iodine alters gene expression in the MCF7 breast cancer cell line: evidence for an anti-estrogen effect of Iodine. *International Journal of Medical Sciences*. 5(4)189–196.