Symbolic Data Analysis to Defy Low Signal-to-Noise Ratio in Microarray Data for Breast Cancer Prognosis

Lyamine Hedjazi, Marie-Véronique V Le Lann, Tatiana Kempowsky-Hamon, Florence Dalenc, Joseph Aguilar-Martin, Gilles Favre

Symbolic Data Analysis to Defy Low Signal-to-Noise Ratio in Microarray Data for Breast Cancer Prognosis

LYAMINE HEDJAZI,[1,2,*] MARIE-VERONIQUE LE LANN,[1,2]
TATIANA KEMPOWSKY,[1] FLORENCE DALENC,[3] JOSEPH AGUILAR-MARTIN,[1]
and GILLES FAVRE[3]

ABSRACT

Microarray profiling has brought recently the hope to gain new insights into breast cancer biology and thereby improve the performance of current prognostic tools. However, it also poses several serious challenges to classical data analysis techniques related to the characteristics of resulted data, mainly high-dimensionality and low signal-to-noise ratio. Despite the tremendous research work performed to handle the first challenge in the feature selection framework, very little attention has been directed to address the second one. We propose in this paper to address both issues simultaneously based on symbolic data analysis capabilities in order to derive more accurate genetic marker-based prognostic models. In particular, interval data representation is employed to model various uncertainties in microarray measurements. A recent feature selection algorithm that handles symbolic interval data is used then to derive a genetic signature. The predictive value of the derived signature is then assessed by following a rigorous experimental setup and compared to existing prognostic approaches in terms of predictive performance and estimated survival probability. It is shown that the derived signature (GenSym) performs significantly better than other prognostic models, including the 70-gene signature, St. Gallen and NIH criterions.

[1]CNRS, LAAS, 7 avenue du Colonel Roche, F-31077 Toulouse, France.
[2] Université de Toulouse, INSA, LAAS, F-31077 Toulouse, France.
[3]Institut Claudius Regaud, Toulouse, F-31052, France.
[*]Corresponding author: lhedjazi@laas.fr, Tel: +33561336947, Fax: +33561336936.

1 INTRODUCTION

Breast cancer management has been for a long time guided by the clinical and histo-pathological knowledge gained from many decades of cancer research. However, the high mortality from breast cancer has pushed researchers to seek for accurate cancer prognosis tools that help physicians to take the necessary treatment decisions that spare patients from side effects and thereby reduce its high medical costs. In the past decade microarray analysis has had a great interest in cancer management such as diagnosis (Ramaswamy et al., 2001), prognosis (Van't Veer et al., 2002), and treatment benefit prediction (Straver et al., 2009). However, the introduction of this technology has brought with it new serious challenges related mainly to high dimensionality of microarray data (or high feature-to-sample ratio) and its low signal-to-noise-ratio.

It has been reported recently that the major difficulty in deciphering high throughput gene expression experiments comes from the noisy nature of the data (Tu et al., 2002). Indeed, data issued from this technology are not only characterized by the dimensionality problem but present also another challenging aspect related to their low signal-to-noise ratio. The noise in such type of data is multisource: biological and noisy measurement, slide manufacturing errors, hybridization errors, scanning errors of hybridized slide (Tu et al., 2002; Nykter et al., 2006). Biological errors are typically due to internal stochastic noise of the cells and error sources related to sample preparation (Blake et al., 2003). This type of intrinsic noise is present in all measurements, regardless of the measurement technology. Measurement errors, on the other hand, include error sources that are a kind of extrinsic noise directly related to the measurement technology and its limitation (e.g. bias due to the used dyes) (Nykter et al., 2006, Blake et al., 2003). Slide manufacturing errors are related to microarray slide images. These include variation in the spot position and size. In addition the marks done by a print tip and deformations in the spot shape can be produced. Hybridization errors include background noise, spot bleeding, scratches, and air bubbles (Nykter et al., 2006).

Appropriate position for Figure 1

Another possible source of error is the digitization of hybridized slide by scanning. The hybridized slide is read by scanning each dye color separately, it might be possible that channels do not align perfectly (Nykter et al., 2006). Many studies were performed to study the different effects of experimental, physiological, and sampling variability (Lee et al., 2000; Novak et al., 2002). An interesting study has been performed in Tu et al. (2002) to analyze the quantitative noise in gene expression microarray experiments. The authors have shown through two illustrative concrete examples the difference in gene expression due to experimental noises. In the first example, a comparison between gene expression values measured on the same sample has been performed. Figure 1 (a) shows the overall difference in two measured gene expression due to measurement error alone as provided in Tu et al. (2002). The deviation of the scattered points from the diagonal line represents the difference between the two measured transcriptomes. In the second example two samples from different cultures are compared as shown in Figure 1 (b) so that the measured expression value differences contain the combined effect of the genuine gene expression differences caused by measurement error.

Although Figures (a) and (b) appear similar, the causes of deviations in the expression values from the diagonal line are completely different. The first one is due only to gene expression measurement error whereas the second is due to the combined effect of the gene expression differentiation and measurement error. Therefore, it is crucial to characterize the difference caused purely by experimental measurement from the expression differentiation due to the difference between the two cultures.

Most of breast cancer studies performed using classical classification and feature selection approaches for microarray data analysis assume that data is perfect without wondering about its reliability. One common practice to deal with this problem is to transform in a non-linear way the gene-expression levels in a preprocessing phase so that the variance across experiments becomes comparable for each gene (Huber et al., 2002). A drawback with this approach is that a global transformation does not adequately account for the fact that the same gene may be measured with different precision in different experiments. Another drawback with this approach is that a complex non-linear transformation of the data complicates measurement interpretation when compared to a global transformation.

We propose here to address this problem within machine learning framework in the aim to design more accurate breast cancer management tools to help the physicians in their decision making process. An interesting approach would be to use symbolic data analysis (SDA) popularized by Bock and Diday (2000). Within this framework, interval data representation can be used to take into account the uncertainty and noise inherent to measurements (Billard, 2008). Symbolic interval features are extensions of pure real data types, in the way that each feature may take an interval of values instead of a single value (Gowda and Diday, 1992). In this framework, the value of a quantity *x* (e.g. gene expression value) is expressed as a closed interval $[x^-,x^+]$ whenever *x* is noised or uncertain; representing the information that $x^- \le x \le x^+$. The uncertainty can be related to the incapability to obtain true values due to possible variability under some changing and complex experimental conditions. However, the introduction of interval representation makes the data processing task more complex than when only a numerical value is considered, especially when high dimensionality problem is faced jointly. Therefore, what is really needed is an approach that enables to process efficiently high dimensional interval datasets. We take advantage here of our recently proposed algorithm (Referred to here as InterSym) that supports such requirements to derive a gene signature for cancer prognosis from microarray datasets.

In the next section we describe how the uncertainties can be integrated in microarray data through the use of interval representation. We give then in section 3 a brief description of the interval feature selection algorithm used here to process the issued interval dataset in order to derive a genetic signature. In section 4 we investigate the proposed strategy on a popular prognostic dataset. We show how the proposed strategy can be used to derive genetic signatures by following a rigorous experimental protocol. The effectiveness of the derived model has been compared with existing prognostic approaches based either on clinical or genetic markers.

2 DATASET

2.1 Raw dataset

The study is performed using the well-known van’t Veer dataset (van’t Veer et al., 2002). van’t Veer and colleagues used a dataset containing 78 sporadic lymph-node-negative patients younger than 55 years of age and less than 5 cm in tumor size, to derive a prognostic signature in their gene expression profiles. Forty-four patients remained disease-free after their initial diagnosis for an interval of at least 5 years (good prognosis group), and 34 patients had developed distant metastases within 5 years (poor prognosis group). We use the same group of patients in the aim to derive a gene prognostic signature. Patient with missing data (1 poor prognosis patient) was excluded in our study. We describe hereafter how this data set is used to generate an interval microarray dataset using the interval representation to model different uncertainties.

2.2 Interval dataset generation

In order to take into account the uncertainty in gene expression measurements under the form of symbolic intervals, an appropriate setup should be followed. Let the m gene expression levels be initially represented in a matrix Y=[$y_1,y_2,...,y_m$] where m is the number of genes. The microarray interval dataset generation is performed by adding a white Gaussian noise with a specific Signal-to-Noise Ratio (SNR=3). Let's consider that the added white Gaussian noise has an absolute value b_j, then the value of the j^{th} interval feature $x_j=[x_j^-, x_j^+]$ corresponding to the j^{th} gene having an expression y_j is obtained as follows:

$$x_j^- = y_j - b_j$$

$$x_j^+ = y_j + b_j$$

It results that

$$x_j = [x_j^-, x_j^+] = [\, y_j - b_j, y_j + b_j].$$

At the end of this step the m gene expression levels are represented in a matrix X=[$x_1,x_2,...,x_m$] where x_j is an interval vector. Once the microarray interval dataset is obtained, a genetic signature can be derived using a feature selection algorithm handling interval data. We use for that our feature selection algorithm proposed recently in Hedjazi et al. (2011), referred to as InterSym, to build a computational model that accurately predicts the risk of distant recurrence after 5-years period of breast cancer diagnosis.

For a better conditioning of magnitudes and processing time minimization, a simple linear re-scaling of raw interval values within the interval [0,1] will also be usually performed:

$$x_i^- = \frac{\hat{x}_i^- - \hat{x}_{i\min}^-}{\hat{x}_{i\max}^+ - \hat{x}_{i\min}^-}, \quad x_i^+ = \frac{\hat{x}_i^+ - \hat{x}_{i\min}^-}{\hat{x}_{i\max}^+ - \hat{x}_{i\min}^-} \tag{1}$$

3 INTERVAL FEATURE SELECTION

The emergence of microarray technology has made possible the simultaneous measurement of the expression of thousands of genes. This technology has carried with it the hope to gain new insights into cancer biology and may improve current tools for cancer management. However, this technology has also brought serious challenges related to intrinsic characteristics of the resulting data. Mainly two challenges are faced simultaneously: (1) high data dimensionality (thousands of gene expressions for a small number of samples); and (2) the noisy nature of measurements (or low signal-to-noise ratio). Since traditional statistical methods are ill-conditioned to deal with such problems, machine learning approaches have been picked up as a good alternative to overcome these difficulties (Haibkains, 2009). The first challenge has been already extensively addressed by using feature selection algorithms. During the past decades, feature selection has indeed played a crucial role in problems involving a huge number of features by selecting only the most relevant features for the problem under investigation. Here, we use the term feature to refer to a gene marker. Existing feature selection algorithms are traditionally characterized as wrappers and filters according to the criterion used to search for the relevant features (Kohavi and John, 1997; Guyon and Elisseeff, 2003). Wrapper algorithms optimize the performance of a specified machine-learning algorithm to assess the usefulness of the selected feature subset; whereas filter algorithms use an independent evaluation function based generally on a measure of information content (entropy, t-test,…) (Kohavi and John, 1997; Guyon and Elisseeff, 2003). Filter algorithms are computationally more efficient but perform worse than wrapper algorithms (Kohavi and John, 1997; Guyon and Elisseeff, 2003). Thereby, with filter algorithms the features are evaluated individually without taking into account the correlation information and redundancy problems. Hence, this can deteriorate drastically the classifier performance (Kohavi and John, 1997). On the other hand, the noisy nature of microarray measurement poses a great challenge for the existing machine-learning algorithms. However, unlike the high-dimensionality problem, a very little attention has been devoted to address this problem by the machine-learning community. Therefore, it is crucial to design efficient feature selection algorithms able to address both problems jointly in order to improve cancer management. One natural idea would be to take use of interval representation to model measurement uncertainty in microarray data. However, this will produce high-dimensional interval datasets which makes the feature selection task even more challenging. Although traditional feature selection algorithms are proficient for processing high-dimensional numerical data, they remain inappropriate for interval data. In the particular case where feature interval values are regular[1], a common practice to apply such algorithms is to label interval values by integers, introducing a metric which is not necessarily the same as in the original data. This can be a potential source of distortion and information loss. In most real applications a feature measurement presents generally a large variation in term of uncertainty and noises from one sample to another, and should be therefore expressed by overlapped intervals. The transformation interval-to-integer in this case is no longer possible and classical algorithms become inapplicable.

We have recently proposed a new interval feature selection algorithm, referred to as InterSym (Hedjazi et al., 2011), which alleviates the previously mentioned problems. InterSym enables to process the interval features in their original form without any restriction on their relative positions (overlapped or regular); no arbitrary mapping is therefore required. To avoid the heuristic search during the feature selection procedure, InterSym optimizes an objective function using classical optimization techniques. The feature's importance is evaluated within a similarity margin framework. Since we address a problem with only two classes (i.e. metastasis or no metastasis), we limit the description of InterSym in this paper for binary class problems.

[1] Interval features take their values from an accountable set of interval values.

Let $D = [x_n, C_k]_{n=1}^{N} \in X \times C$ be the training dataset, where x$_n$=[$x_n^1, x_n^2, \ldots, x_n^m$] is the *n*-th data sample containing *m* features, C$_k$ its corresponding class label, and x_n^i stands for the i^{th} interval value included in its domain U^i. The first step of InterSym algorithm concerns the parameterization of each class by an interval vector based on an appropriate learning process through the following arithmetic means:

$$\rho_k^{i-} = \frac{1}{N_k} \sum_{j=1}^{N_k} x_j^{i-} \text{ ,and } \rho_k^{i+} = \frac{1}{N_k} \sum_{j=1}^{N_k} x_j^{i+} \qquad (2)$$

The resulted class prototype for all the features is given by $\rho_k = \left[\rho_k^1, \rho_k^2, \ldots, \rho_k^m\right]^T$ where $\rho_k^i = \left[\rho_k^{i-}, \rho_k^{i+}\right]$. Then, a similarity measure has been defined in Hedjazi et al. (2011) to estimate the feature resemblance of the i^{th} interval feature value $x_n^i = \left[x_n^{i-}, x_n^{i+}\right]$ of sample x$_n$ to each class represented by its interval prototype $\rho_k^i = \left[\rho_k^{i-}, \rho_k^{i+}\right]$:

$$S\left(x_n^i, \rho_k^i\right) = \frac{1}{2}\left(\frac{\varpi\left[x_n^i \cap \rho_k^i\right]}{\varpi\left[x_n^i \cup \rho_k^i\right]} + 1 - \frac{\partial\left[x_n^i, \rho_k^i\right]}{\varpi\left[U^i\right]}\right) \qquad (3)$$

Where $\varpi[I] = \left|I^- - I^+\right|$ and $\partial\left[x_n^i, \rho_k^i\right] = \max\left[0, (\max\left[x_n^{i-}, \rho_k^{i-}\right] - \min\left[x_n^{i+}, \rho_k^{i+}\right]\right]$.

U^i states for the domain of i^{th} interval feature values.

We assume that the n^{th} data sample x$_n$=[$x_n^1, x_n^2, \ldots, x_n^m$] is labeled by class c. Let $\tilde{c}$ be the alternative class. Based on the similarity measure (3), two similarity vectors can be associated to each data sample as follows

$$\Gamma_{nc} = \left[S\left(x_n^1, \rho_c^1\right), S\left(x_n^2, \rho_c^2\right), \ldots, S\left(x_n^m, \rho_c^m\right)\right]^T$$

$$\Gamma_{n\tilde{c}} = \left[S\left(x_n^1, \rho_{\tilde{c}}^1\right), S\left(x_n^2, \rho_{\tilde{c}}^2\right), \ldots, S\left(x_n^m, \rho_{\tilde{c}}^m\right)\right]^T \qquad (4)$$

A similarity margin for sample x$_n$ can be defined as

$$\vartheta_{nc} = \phi(\Gamma_{nc}) - \phi(\Gamma_{n\tilde{c}}) \quad (5)$$

where Γ_{nc} and $\Gamma_{n\tilde{c}}$ are respectively the similarity vectors of sample x_n to classes c and $\tilde{c}$, $\phi(y) = 1/m \sum_{i=1}^{m} y_i$ is a function representing the global similarity of the sample x_n to the given class.

A weighted similarity margin can be defined through a weight assignment in the previously defined similarity margin to express the importance of each interval feature as follows

$$\vartheta_{nc}(\mathrm{w}) = \phi(\Gamma_{nc} / \mathrm{w}) - \phi(\Gamma_{n\tilde{c}} / \mathrm{w}) = \frac{1}{m} . \sum_{i=1}^{m} w_i . \left(S\left(x_n^i, \rho_c^i\right) - S\left(x_n^i, \rho_{\tilde{c}}^i\right) \right) \quad (6)$$

Note that a sample x_n is considered correctly classified if $\vartheta_{nc} \succ 0$. A natural idea to estimate the interval feature weight is to maximize the leave-one-out classification error as follows:

$$\begin{aligned} &\underset{w}{Max} \sum_{n=1}^{N} \vartheta_{nc}(\mathrm{w}) = \underset{w}{Max} \frac{1}{m} \sum_{n=1}^{N} \left(\sum_{i=1}^{m} w_i . \left(S\left(x_n^i, \rho_c^i\right) - S\left(x_n^i, \rho_{\tilde{c}}^i\right) \right) \right) \\ &s.t. \ \|\mathrm{w}\|^2 = 1, \mathrm{w} \geq 0 \end{aligned} \quad (7)$$

Where ϑ_{nc} is the margin of x_n computed with respect to the weight vector w. The first constraint is the normalized bound for the modulus of w so that the maximization ends up with non infinite values, whereas the second guarantees the nonnegative property of the obtained weight vector. A closed-form solution can be obtained using the classical Lagrangian optimization approach:

$$\mathrm{w}^* = \frac{\mathrm{r}^+}{\|\mathrm{r}^+\|}$$

$$\mathrm{r} = \frac{1}{m} \sum_{n=1}^{N} \{\Gamma_{nc} - \Gamma_{n\tilde{c}}\} \quad (8)$$

With $\mathrm{r}^+ = [\max(r_1,0), \ldots, \max(r_m,0)]^T$

InterSym is considered as one of the first feature selection algorithms that enable processing interval feature-type data. Note that the objective function optimized by InterSym approximates the leave-one-out cross validation error and thus chooses only the features if they contribute to the overall performance. Hence, both issues, correlation and redundancy, are addressed by InterSym. Moreover, InterSym avoids the heuristic combinatorial search by using classical optimization approaches to achieve an analytical solution. Furthermore, an extension of InterSym has been also proposed for multiclass problems (Hedjazi et al., 2011). The effectiveness of InterSym in (Hedjazi et al., 2011) has been shown through three real-world applications on low-dimensional interval datasets. However, it would be interesting to assess its effectiveness also on high dimensional problems such as microarray interval datasets. Subsquently, we apply InterSym algorithm to derive a genetic signature for breast cancer prognosis, by taking into account the measurement uncertainty through the use of interval representation. As mentioned previously, InterSym will enable the selection of relevant information in high-diemensional interval datasets by avoiding any related numerical and heuristic search complexities.

4 EXPERIMENTS AND RESULTS

4.1 Experimental setup

Data issued by microarray technology provides the measurement of thousands of gene expressions for usually small number of patients. This situation can likely lead to serious problem of overfitting of the computational model on training data, i.e. the model performs very well on training data while achieve extremely poor results on unseen data. A special experimental protocol therefore is generally adopted to avoid this problem such as cross-validation protocols. Due to the small sample size in our case we performed a LOOCV (Leave One-Out Cross Validation) to estimate the optimal classification parameters as proposed in (Wessels et al., 2005). In each iteration of this procedure, one sample is held-out for testing and the remaining samples are used for training. The training data are used to estimate the optimal parameters of the classifier and to perform the feature selection task. The resulting model is employed then to classify the held-out sample. This experiment is carried out on all samples so that each of them has been used once for testing.

Very few classification methods are capable to deal with interval representation particularly if intervals may overlap. Therefore, we choose to use here LAMDA classifier (Learning Algorithm for Multivariate Data Analysis) (Hedjazi et al., 2012), able to handle efficiently interval data as well as numerical and qualitative data, to demonstrate the predictive values of the derived prognostic signature by InterSym and comparing its performance with those of existing approaches such as clinical-based approaches (St-Gallen, all clinical markers,…) and genetic-based approaches (70-gene signature). For this classifier only one parameter needs to be specified in the training phase (exigency index).

It is worthwhile to note here that in the study performed by van't Veer and colleagues, a 70-gene signature has been derived from the same dataset using a feature selection method based on correlation coefficient. The predictive value of the 70-gene has been then assessed by using a correlation based classifier (van't Veer et al., 2002).

4.2 Results

A genetic signature, referred to here as GenSym, was derived based on the InterSym algorithm corresponding to the optimal classification performance using the LAMDA classifier. We note that both of InterSym and LAMDA enable to handle appropriately interval data for classification and feature selection respectively (see previous sections for more details). Table 1 shows the classification performance obtained with LAMDA using GenSym signature. For comparison, classification performance using 70-gene signature, clinical markers, St-Gallen consensus and NIH criterion are also reported in Table 1. We observe that the GenSym signature significantly outperforms the 70-gene, clinical and classical clinical criterions (St-Gellen, NIH).

Appropriate position for Table 1

GenSym achieves indeed a high accuracy (~90%) while significantly improves specificity and sensitivety of the 70-gene signature (by more than 6 % and 10% respectively). It should be noted also that in the study performed by van't Veer and colleagues the sensitivity level has been set to 90% in order to ensure a high classification rate of poor prognosis patients, which has led to a poor specificity level (72%). GenSym, however, while providing a sensitivity level close to the threshold imposed by van't Veer and colleagues, it ensures a similar high level of specificity enabling therefore to spare a big number of good prognosis patients from receiving unnecessary toxic treatment.

Classification performance is not always a sufficient criterion for comparing predictive values of different marker signatures. Performance measurement can also depend strongly on a decision threshold when only a limited number of patients are available. Varying this decision threshold enables to visualize the performance of a given classifier over all sensitivity and specificity levels through a Receiver Operating Characteristic (ROC) curve.

For further comparisons of the different approaches, we plotted in Figure 2 the ROC curve for GenSym, 70-genes and clinical-based approaches. The St-Gallen and NIH criteria are not shown here since the good prognosis group contains very few patients. It can be observed that the GenSym signature significantly outperforms the 70-gene signatures as well as clinical markers over almost all sensitivity and specificity ranges.

Appropriate position for Figure 2

We performed also survival data analysis of the four approaches, GenSym signature, 70-gene signature, clinical markers and St-Gallen criterion, to further demonstrate the prognostic value of the GenSym signature. The Kaplan-Meier curves with 95% confidence intervals of respectively the four approaches are shown in Figure 3. Particularly the GenSym signature induces a significant difference in the probability of remaining metastases-free in patients with a good signature and the patients with a poor prognostic signature (P-value<0.001). Hazard Ratio (HR) estimated by Mantel-Cox approach of distant metastases within five years for the GenSym signature is 8.20 (95% CI: 4.16- 16.2), which is superior to either the 70-gene signature, St Gallen consensus or clinical markers. The HR obtained for clinical and St Gallen consensus (respectively 2.32 (95% CI: 1.36- 3.95) and 1.17 (95% CI: 0.46- 2.92)) are consistent with those reported in many similar studies (Wang et al., 2005; Soutiriou et al., 2006), suggesting that clinical markers have bad predicting value.

Appropriate position for Figure 3

4.3 GenSym signature

The GenSym signature is composed from 23 genes, given in Table 2, among them 12 genes are listed in the 70-gene signature. Although the held-out testing sample is not involved in the identification of a gene signature in each iteration, it should be noted that we have find that the identified signature stays relatively stable over all iterations. The functional annotation for the genes should provide insight into the underlying biological mechanism leading to rapid metastases. According to the National Center for Biotechnology Information (NCBI) databases, among the GenSym signature, genes involved in proliferation, invasion and metastasis are significantly unregulated in the metastasis group. For instance we find TSPYL5 which has been revealed to play important roles in modulation of cell growth and cellular response probably via regulation of the akt signaling pathway. It is reported that TSPYL5 is a poor prognosis marker and reduces the p53 protein levels and inhibits activation of p53-target genes. It is worthwhile to note here that this is the top listed gene in the 70-gene signature. MMP-9 is also related to tumor invasion and metastasis by their capacity for tissue remodeling via extracellular matrix as well as basement membrane degradation and induction of angiogenesis. Evaluation of MMP-9 expression seems to add valuable information on breast cancer prognosis.

Appropriate position for Table 2

GenSym signature holds also many new meaningful genes (such as FBP1, IGFBP5, FGF18, SSX1, NUSAP1, C1GALT1, BTG2, PEX12). The importance of both (FBP1, IGFBP5) can be highlighted by the actually suspected relation between the insulin and tumor growth (Becker et al., 2012). But neither FBP1 nor IGFBP5 have been evaluated independently in human cancers. However, FBP1 has been also found strongly associated with disease outcome among the 231 top ranked genes in (van't Veer et al., 2002). FGF18 has been revealed clearly involved in the carcinogenesis of ~10% breast cancer. NUSAP1 has also been found to be related to proliferation and cells division. SSX1 is involved in certain sarcomas; it controls the cell cycle and is considered as an important transcription factor. C1GALT1 is a protein that plays an important role in cell adhesion whereas BTG2 is considered as a tumor suppressor.

5 CONCLUSION

In this paper, we addressed the problem of low signal-to-noise ratio in microarray data faced jointly with high data dimensionality problem. The basic idea is to take advantage of symbolic data analysis capabilities to alleviate this issue, suggesting the use of interval representation to model uncertainty in microarray measurements. We derived then based on our recently proposed interval feature selection algorithm a genetic signature. The GenSym signature holds some common genes with existing genetic signatures as well as new genes showing a meaningful biological interpretation and high relevance to the biology of breast cancer disease. We have shown through a preliminary computational study that the use of such strategy can improve and simplify significantly the cancer prognosis task by selecting a small number of relevant genetic markers as compared to other existing signatures (only 23 markers). Its predictive value has been assessed also through this study and compared with existing genetic signatures and clinical criterions. We believe that the proposed strategy will open the door wide for the introduction of a new generation of symbolic algorithms in bioinformatics applications.

To further demonstrate the effectiveness of the proposed strategy, a larger-scale experimental study is now underway in the framework of a research project using a large number of patients issued from publicly available datasets.

ACKNOWLEDGEMENTS

Funding: This work was supported in part by the ANR project INNODIAG involving the companies Den Dendris, Inno Innopsys and laboratories LAAS, ICR, ITAV, PFBT.

REFERENCES

Becker, M.A., Hou, X., Harrington, S.C., et al. 2012. IGFBP ratio confers resistance to IGF targeting and correlates with increased invasion and poor outcome in breast tumors. *Clin. Cancer Res.,* doi: 10.1158/1078-0432.CCR-11-1806.

Billard, L. 2008. Somes Analyses of Interval Data. *J of Comp and Info Tech*. **16**, 225-233.

Blake, W.J., KÆrn, M., Cantor, C.R., et al. 2003. Noise in eukaryotic gene expression. *Nature*. **422** (6932), 633-637.

Bock, H.H. and Diday,E. 2000. Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data. Springer-Verlag, Heidelberg.

Gowda, K.C. and Diday,E. 1992. Symbolic clustering using a new similarity measure. *IEEE Trans. Systems Man Cybernet*. **22**, 368-378.

Guyon,I. and Elisseeff,A. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res*. **3**, 1157-1182.

Haibe-Kains,M.B. 2009. Identification and Assessment of Gene Signatures in Human Breast Cancer [Ph.D. dissertation]. Université Libre de Bruxelles.

Hedjazi,L., Aguilar-Martin, J., Le Lann, M-V., et al. 2011. Similarity-margin based feature selection for symbolic interval data. *Pattern Recognition Letters*. **32 (4)**, 578-585.

Hedjazi,L., Aguilar-Martin, J., Le Lann, M-V., Kempowsky, T., et al. 2012. Towards a Unified Principle for Reasoning about Heterogeneous Data: A Fuzzy Logic Framework, *Int'l J. of Uncertainty, Fuzziness and Knowledge-Based Systems. 20 (2), 281-302*.

Huber,W., Heydebreck, Sültmann, H., et al. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, 96-104.

Kohavi,R. and John,G.H. 1997. Wrapper for feature subset selection. *Artificial Intelligence*. **97**, 273-324.

Lee, M-L.T., Kuo, F.C., Whitmore, G. A., et al. 2000. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA.* ***97***, pp. 9834-9839.

Novak, J.P., Sladek, R., Hudson, T.J., et al. 2002. Characterization of variability in large-scale gene expression data: implications for study design. *Genomics*. **79**, 104-113.

Nykter, M., Aho, T., Ahdesmäki, M., et al. 2006. Simulation of microarray data with realistic characteristics. *BMC Bioinformatics*. **7**(1), 332-349.

Ramaswamy, S., Tamayo, P., Rifkin, R., et al. 2001. MultiClass Cancer Diagnosis Using Tumor Gene Expression Signatures. *Proc. Nat'l Acad. Sc. USA.* ***98*** **(26)** , 15149-15154.

Sotiriou, C., Wirapati, P., Loi, S., et al. 2006. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst.* **98(4)**, 262-72.

Straver, M.E., Glas, A.M., Hannemann, J., et al. 2009. The 70-gene signature as a response predictor for neoadjuvant chemotherapy in breast cancer. *Breast Cancer Res. Treat.* **119**, 551-558.

Tu, Y., Stolovitzky, G., Klein, U., et al. 2002. Quantitative noise analysis for gene expression microarray experiments. *Proc. Nat'l Acad. Sc. USA*. **99** (22), 14031-14036.

van't Veer, L.J., Dai, H., van de Vijver, M.J., et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* **415**, 530-536.

Wessels, L.F.A., Reinders, M.J.T, Hart, A.A.M., et al. 2000. A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*. **21**, 3755-3762.

Wang, Y., Klijn, J.G., Zhang, Y., et al. 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. **365 (9460)**, 671-679.

Table 1. Comparatives results between 23-gene signature and existing approaches

Method	TP	FP	FN	TN	Sens.	Spec.	Acc.
GenSym	29/33	4/44	4/33	40/44	**87.88**	**90.91**	**89.61**
70-gene[(a)]	27/33	9/44	6/33	35/44	81.82	79.55	80.52
Clinical	26/33	14/44	7/33	30/44	78.79	68.18	72.73
St-Gallen[(b)]	33/33	39/44	0/33	5/44	100	6.49	50.65
NIH[(c)]	33/33	44/44	0/33	0/44	100	0	42.86

TP: True Positive**; FP:** False Positive**; FN:** False Negative**; TN:** True Negative**; Sens.:** Sensitivity**; Sens.**: Sensitivity; **Spec:** Specificity.

(a) : Mammaprint signature ©

(b): St. Gallen - Chemio when one criteria is satisfied: ER negative; Lymph node positive; T>2cm; Grade III or II; Age <35 years.

(c): NIH: Chemio when Lymph Node positive or Tumour size > 1cm

Table 2. List of genes included in GenSym and their notations

Rank	Gene ID	70-gene	Notation
1	Contig37063_RC	□	N/A
2	Contig26388_RC	□	N/A
3	NM_003748	■	ALDH4A1
4	NM_006681	■	NMU
5	NM_000507	□	FBP1
6	AF055033	■	IGFBP5
7	NM_000286	□	PEX12
8	AL080059	■	TSPYL5
9	Contig33814_RC	□	N/A
10	NM_012429	□	SEC14L2
11	NM_000599	■	IGFBP5
12	NM_003862	■	FGF18
13	Contig63649_RC	■	N/A
14	NM_004994	■	MMP9
15	Contig11065_RC	■	N/A
16	Contig32185_RC	■	N/A
17	NM_016359	■	NUSAP1
18	Contig15954_RC	□	N/A
19	NM_005635	□	SSX1
20	Contig49388_RC	■	N/A
21	Contig52554_RC	□	N/A
22	NM_020156	□	C1GALT1
23	NM_006763	□	BTG2

■: Listed in 70-gene signature, □: Not listed in 70-gene signature, N/A: Not Available

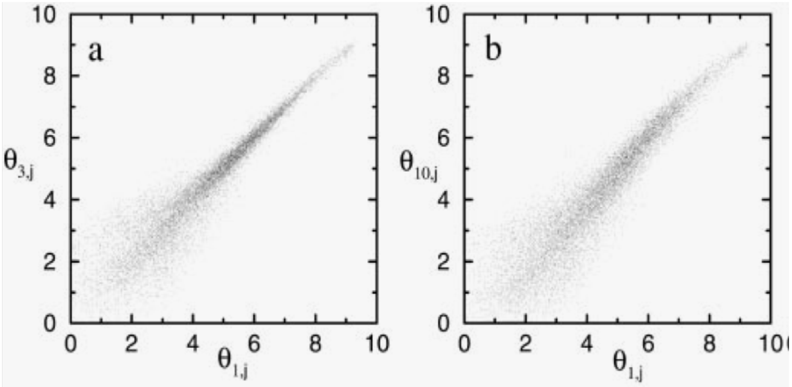

Fig. 1. The scatter plot of gene expression pairs (a) experiments pair on the same sample (b) experiment pair between two different samples. Figure taken from (Tu et al., 2002).

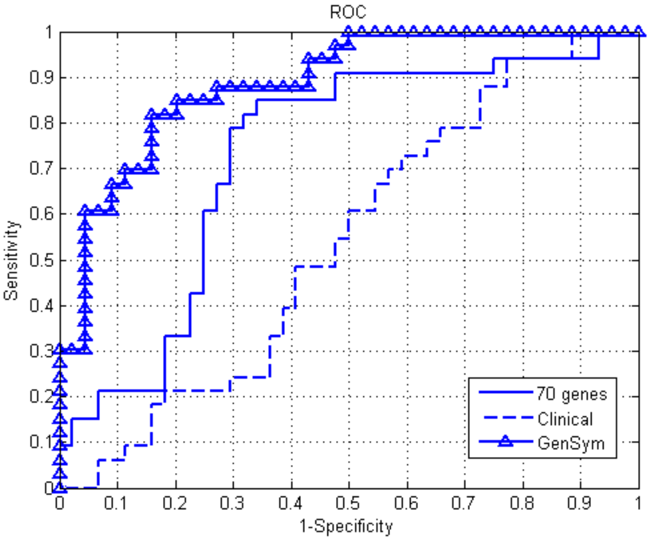

Fig. 2. ROC curve of GenSym, 70-gene and clinical approaches

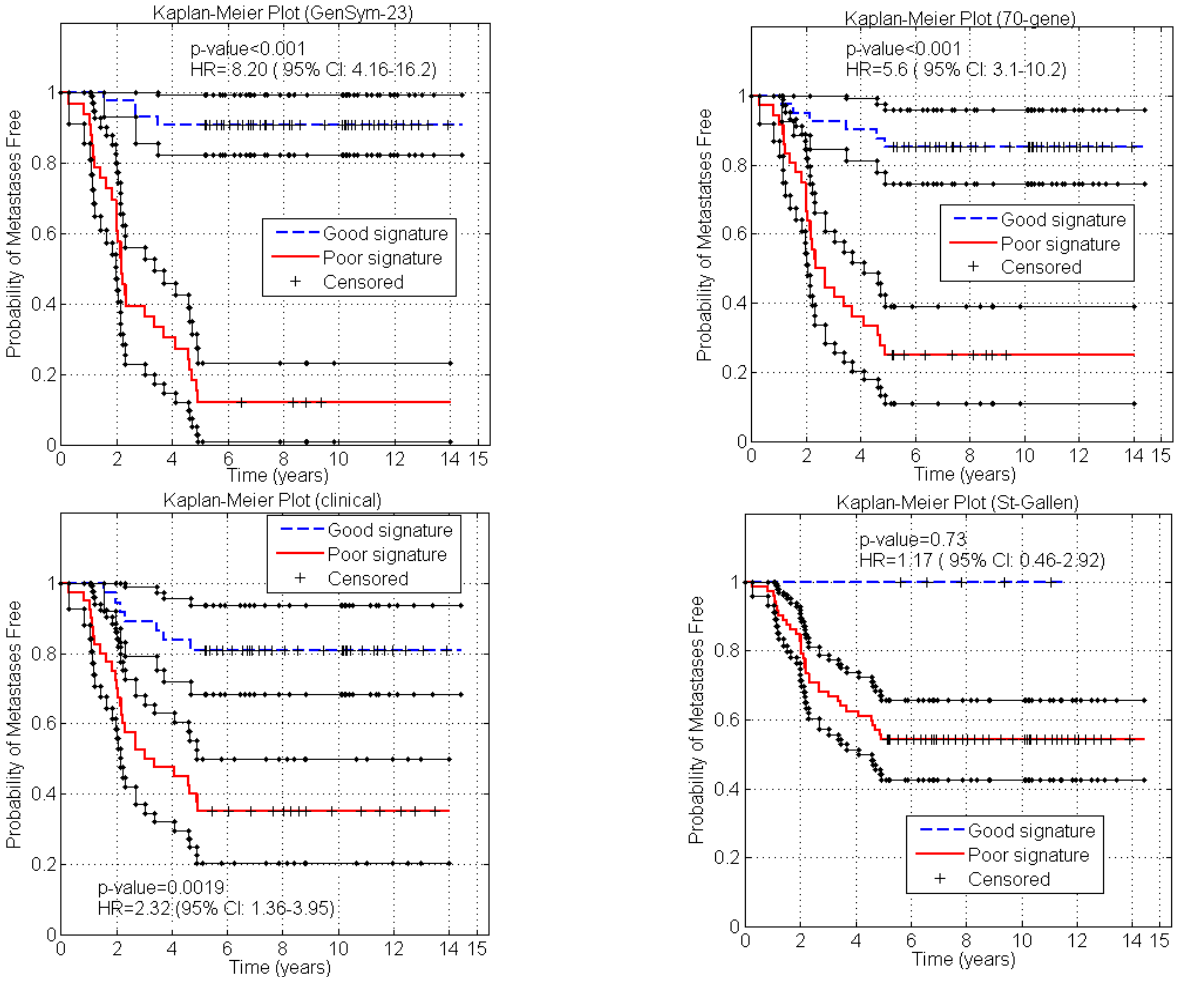

Fig. 3. Kaplan-Meier estimation of the probabilities of remaining metastases-free for the good and poor prognosis groups. The p-value is computed by using log-rank test.