Collateral Missing Value Estimation: Robust Missing Value Estimation for Consequent Microarray Data Processing

Muhammad Shoaib B. Sehgal, Iqbal Gondal, and Laurence Dooley

Faculty of IT, Monash University, Churchill VIC 3842, Australia
{Shoaib.Sehgal, Iqbal.Gondal, Laurence.Dooley}
@infotech.monash.edu.au

Abstract. Microarrays have unique ability to probe thousands of genes at a time that makes it a useful tool for variety of applications, ranging from diagnosis to drug discovery. However, data generated by microarrays often contains multiple missing gene expressions that affect the subsequent analysis, as most of the times these missing values are ignored. In this paper we have analyzed how accurate estimation of missing values can lead to better subsequent gene selection and class prediction. *Collateral Missing Values Estimation* (CMVE), which demonstrates superior imputation performance compared to *Bayesian Principal Component Analysis* (BPCA) Impute, *K-Nearest Neighbour* (KNN) algorithm, when estimating missing values in the BRCA1, BRCA2 and Sporadic genetic mutation samples present in ovarian cancer by exploiting both local/global and positive/negative correlation values. CMVE also consistently outperforms, in terms of classification accuracies, BPCA, KNN and *ZeroImpute* techniques. The imputation is followed by gene selection using fusion of *Between Group to within Group Sum of Squares* and *Weighted Partial Least Squares* where *Ridge Partial Least Square* algorithm is used as a class predictor.

1 Introduction

Microarrays have has wide range of applicability from diagnosis to drug discovery due to their ability to probe tens or thousands genes at a time [1, 2]. Despite this however, microarray data frequently contains missing values due to spotting problems, slide scratches, blemishes on the chip, hybridization error, image corruption or simply dust on the slide [3]. These missing values affect subsequent inference from: gene selection, class prediction and data dimension reducing techniques such as *Between Group to within Group Sum of Squares* (BSS/WSS) [4], *Neural Networks* (NN), *Support Vector Machines* (SVM), *Principal Component Analysis* (PCA) and *Singular Value Decomposition* (SVD) [5, 6]. Different strategies to solve the problem of missing data can be adopted. The simplest method is to repeat the process, though this is seldom feasible for economic reasons or ignoring those samples, containing missing values, though this again is not recommended due to limited number of samples available. The best strategy is to attempt to accurately estimate the missing values. Normally missing values are replaced with zero values which doesn't take advantage

S. Zhang and R. Jarvis (Eds.): AI 2005, LNAI 3809, pp. 274 – 283, 2005.

of data correlations, so leading to errors in the subsequent analysis [7]. However, if the correlation between data is exploited then the missing value prediction error can be significantly reduced [8].

Besides this, the number of samples m in microarray data is relatively much less than the number of genes n per sample (usually in thousands) that makes most of the classical class prediction methods to perform poorly and they overfit to the training data [9]. For example, FLD function in *Fisher Linear Discriminant* (FLD) Analysis is singular when $m < n + 2$ [10]. In spite of this if the genes are included for class prediction by classifiers; it includes the associated noise of the data resulting in lower prediction accuracy. This problem can be solved if feature selection is applied to the data.

Most feature selection algorithms are dimension reduction techniques, for example PCA [1] and SVD, do not consider class discrimination while converting data to Eigen space resulting in lower class prediction accuracy. Alternatively, univariate algorithms are used, for example; t-test, signal to noise ratio [11], BSS/WSS [4], *Significance Analysis of Microarray* (SAM) [12] which are either made for binary class response or they consider each relevant gene individually which selects the genes which are highly correlated which it introduces redundancy [13]. The problem can be avoided if multivariate gene selection is applied to simultaneously consider multiple genes and class information, hence reducing redundancy of covariate genes and keeping the class discrimination intact. However, if multivariate method is coupled with class prediction accuracy then it is highly dependent on learning method [13]. Therefore, a suitable strategy is required which can predict the missing values and also can minimize the problems in feature selection techniques.

In this paper we have proposed an innovative solution to the aforementioned problems by applying the recently introduced *Collateral Missing Value Estimation* (CMVE) algorithm [14] that not only guarantees lower prediction error than *Bayesian PCA* (BPCA), *ZeroImpute* and *K Nearest Neighbour* (KNN), but has also increased the classification accuracy for the range of missing values from 1-20% for multiclass ovarian cancer data [15]. To select significant genes, a two fold strategy is applied which uses both univariate and multivariate methods by stacking both algorithms. The p discriminant genes are first selected by univariate BSS/WSS to gain the advantage of model independence and then redundant genes are removed by *Weighted Partial Least Square Method* (WPLS). The other benefit of applying BSS/WSS prior to WPLS is that it reduces computational time by selecting a smaller search space for WPLS. For classification *Ridge Partial Least Squares* (RPLS) is applied by regressing significant genes with ridge penalty [16]. The motivation to employ RPLS came from its better prediction ability than other classification algorithms for multi-class microarray data.

The rest of the paper is organized as follows: Section 2 briefly presents methods for Imputation, Gene Selection and Classification used in this paper. Section 3 analyzes empirical results while conclusions are drawn is Section 4.

2 Review of Gene Selection, Missing Value and Classification Algorithms

The following convention is adopted throughout this paper to present Imputation, Gene Selection and Classification techniques. $Y \in \mathbb{R}^{m \times n}$ is assumed to be the gene expression matrix, where *m* is the number of genes and *n* is the number of samples. In *Y*, every gene *I* is represented by $g_I \in Y$, so *Y* in *n* experiments is organized as:-

$$Y = \begin{bmatrix} g_{1^T} \\ \cdot \\ \cdot \\ \cdot \\ g_{m^T} \end{bmatrix} \in \mathbb{R}^{m \times n}. \tag{1}$$

A missing value in gene *I* for sample *J* is expressed as:-

$$Y(I, J) = g_I(J) = \Xi. \tag{2}$$

Following three Sections outline the Collateral Missing Value Estimation, BSS/WSS Gene Selection and RPLS Class Prediction techniques.

2.1 Collateral Missing Value Estimation

Collateral Missing Value Estimation (CMVE) algorithm estimates missing values using multiple estimation matrices with *Least Square Regression* (LS) , *Non Negative LS* and *Linear Programming* (LP), by regressing *k-ranked covariate* genes $\partial \in \mathbb{R}^{k \times n}$. CMVE imputes missing values by merging three estimation matrices Φ_1 , Φ_2 and Φ_3 , computed using LS, NNLS and LP. To estimate Φ_1 for g_{IJ}, LS regression method [17] is used. LS regression problem for $Y \in \mathbb{R}^{m \times n}$ be expressed as:-

$$\Phi_1 = \zeta + \rho Y + \xi, \tag{3}$$

where ξ is the error term that minimizes the variance in the LS model, ς and ρ are unknown coefficients obtained by minimizing least square error. To estimate Φ_2 and Φ_3 CMVE finds a linear combination of models that best fit ∂ and g_I using NNLS algorithm such as:-

$$\Phi_2 = \sum_{i=1}^{k} \phi + \eta - \sum_{i=1}^{k} \xi^{2}, \tag{4}$$

$$\Phi_1 = \frac{\sum_{i=1}^{k} (\phi^T \times g_I)}{k} + \eta, \tag{5}$$

where ϕ is the vector that minimizes ξ_0 in (6), η is the normal residual and ξ is the actual residual. The objective function in NNLS minimizes, using linear programming techniques, the prediction error ξ_0 so that:-

$$\xi, \phi, \eta = \min(\xi_0), \tag{6}$$

that is $\min(\xi_0)$ is a function that locates the normal vector ϕ with minimum prediction error ξ_0 and residual η. The value of ξ_0 in (6) is obtained from:-

$$\xi_0 = \max(SV(\beta.\phi - g_I{}^T)), \tag{7}$$

where *SV* are the singular values of the difference vector between the dot product β and prediction coefficients ϕ with the gene expression $g_1{}^T$. The final estimate χ for Ξ is formed using:-

$$\chi = \Upsilon.\Phi_1 + \Delta.\Phi_2 + \Lambda.\Phi_3, \tag{8}$$

where probabilities where $\Upsilon = \Delta = \Lambda = 0.33$ ensures an equal weighting to the respective estimates Φ_1, Φ_2 and Φ_3. The rationale for this choice is that as each estimate is highly data dependent, it avoids any bias towards one particular estimate [8]. CMVE derives its superior imputation performance over BPCA and KNN by considering both local/global and positive correlations [14], coupled together with a unique self-correcting error property, which guards against the danger of a wildly initial predictions of the missing values [14].

2.2 Between Group to Within Group Sum of Squares

This gene selection method identifies those genes which have large inter-class variations while concomitantly having small intra-class variations. For any gene *I* in $Y \in \mathbb{R}^{m \times n}$ BSS/WSS is calculated as follows:-

$$BSS(I)/WSS(I) = \frac{\sum_{t=1}^{T}\sum_{q=1}^{Q} F(L_t = q)(\bar{Y}_{qI} - \bar{Y}_I)^2}{\sum_{t=1}^{T}\sum_{q=1}^{Q} F(L_t = q)(Y_{It} - \bar{Y}_{qI})^2}, \tag{9}$$

where *T* is the size of a training sample, *Q* is the number of classes *and F(•)* is a Boolean function which results in 1 if the condition is true, and zero otherwise, $\bar{Y}_J$ denotes the average expression level of gene *I* across all samples and $\bar{Y}_{qI}$ is the average expression level of gene *I* across all samples belonging to class *q*. The genes *G* are ranked from highest to lowest BSS/WSS ratios to form significant gene expression matrix ϑ. The first *p* genes are then selected from ϑ for subsequent class prediction. It is followed by Weighted Partial Least Square (WPLS) to eliminate correlated genes from *p*. The motivation to select genes using BSS/WSS and then WPLS is that BSS/WSS does not select multiple genes simultaneously and hence account for dependency between the genes. Also, BSS/WSS ignores the model uncertainty by predicting set of relevant genes and then predicting relevant class [18]. WPLS accounts for model uncertainty by considering class prediction accuracy. However, if only WPLS is used then selected genes are highly dependent on prediction model [13]. Another reason for employing BSS/WSS is that the gene to sample ratio is reduced, so resulting in a shorter convergence time for WPLS.

2.3 Ridge Partial Least Squares

Ridge Partial Least Squares (RPLS) method uses *Partial Least Squares* (PLS) with the *Penalized Logistic Regression* (PLR) for class prediction. To apply PLS for class prediction, class labels are replace by a pseudo-response variable that has expected value in linear relationship with the covariates because PLS can only handle continuous responses. Therefore, in order to extend PLS to Generalized Linear Models, RPLS replaces pseudo-response variable Z^{∞} at the convergence of *Iterative Reweighted Least Square* (RIRLS) algorithm with ridge penalty. The other advantage of choosing Z^{∞} is that this allows the combination of a regularization and dimension-reduction step. RPLS comprises of three major steps:-

1- Pseudo-response variable Z^{∞} and weighted matrix W^{∞} are computed using:-

$$(Z^{\infty}, W^{\infty}) = RIRLS(L, Y, \lambda), \tag{10}$$

where λ is some positive real constant which is calculated by minimizing the *Bayesian Information Criterion* (BIC) [19] and L is a set containing discrete class labels.

2- Matrices Z^{∞} and W^{∞} are then used to compute $\hat{\alpha} \in \mathbb{R}^{p+1}$ by WPLS method using:-

$$\hat{\alpha}^{PLS,\kappa} = WPLS(Z^{\infty}, Y, W^{\infty}, \kappa), \tag{11}$$

where Y is the input matrix and κ is a positive integer which determines number of iterations.

3- Finally, class response is determined using *Linear Logistic Discrimination* (LLD). In LLD the conditional class probability of response L for a given data Y is:-

$$P(L = 1 | Y = y; \hat{\alpha}), \tag{12}$$

where parameter $\hat{\alpha} \in \mathbb{R}^{p+1}$ is estimated using (11) and p are number of predictor genes determined using BSS/WSS (Section 2.2). The probability P in (12) is computed using:-

$$P(L = 1 | Y = y; \hat{\alpha}) = h([1 \;\; y]\hat{\alpha}), \tag{13}$$

where $h(\eta) = 1/[1+\exp(-\eta)]$. The quantity $h([1 \;\; y]\hat{\alpha})$ is a linear predictor. The log-likelihood of the observations for the parameter $\hat{\alpha}$ is given by:-

$$\hat{l}(\alpha) = \sum_{i=1}^{n} \{L_i \upsilon_i(\hat{\alpha}) - \ln[1 + \exp(\upsilon_i(\hat{\alpha}))]\}, \tag{14}$$

which for all $1 \le i \le n, \upsilon_i(\hat{\alpha}) = (Z\hat{\alpha})_i$ and Z = [Υ_n Y] of size $n \times (p+1)$ and Υ_n is the column matrix of size n. The class label L is 1 if $\wp > 1-\wp$ and zero otherwise where

$$\wp = h([1 \;\; y]\hat{\alpha}). \tag{15}$$

3 Results Analysis

Well tested, ovarian cancer microarray data [20] was used as a test data in all the experiments. The motivation to use this data set is that cancer data contains up/down regulated genes and hence are difficult to predict using estimation algorithms [5]. The data set contained 18, 16 and 27 samples of BRCA1, BRCA2 and sporadic mutations (neither BRCA1 nor BRCA2) respectively. Each data sample contained logarithmic microarray data of 6445 genes. To quantitatively evaluate the performance of CMVE imputation technique, the following *imputation error* and *classification error* estimation measures were employed.

3.1 Imputation Error Measure

For the comparison of different imputation techniques, between 1% and 20% of the values were randomly removed from the BRCA1, BRCA2 and Sporadic dataset samples and the *Normalized Root Mean Square* (NRMS) imputation error ξ computed as:-

$$\xi = \frac{RMS(Y - Y_{est})}{RMS(Y)}, \tag{16}$$

where Y is the original data matrix and Y_{est} is the estimated matrix using either CMVE, BPCA or KNN. The advantage of using (16) for error estimation is that ξ=1 for zero imputation [5].

Different values of k were tested for both KNN and CMVE, with k =10 exhibiting the best results. The plots in Fig. 1(a-e) show the NRMS error in estimating randomly introduced missing values from 1% to 20% for BRCA1, BRCA2 and Sporadic datasets. The results confirm that CMVE performed better than BPCA and KNN (see Fig. 1(a-e)). It is also obvious from the graphs that CMVE exhibited improved robustness at higher missing values, with the reason for these improvements being traced back to the reason explained in Section 2.1, that CMVE exploits the relationship between gene expression values more effectively than BPCA and KNN by considering both global and local, as well as positive and negative data correlations.

3.2 Classification Error Measure

Missing values inevitably affect classification accuracy and gene selection, yet many classifiers only use zero imputation [8]. Our cross validation results show that with the proper estimation of missing values, the gene selection and classification accuracy can be significantly improved [12, 14]. So, for the proof of concept, an alternative way is to test imputation methods by randomly removing values from the data and testing the impact on decision making techniques such as gene selections and classification.

The estimation results in Fig. 2(a-f) confirm that CMVE consistently perform better than BPCA, KNN and *ZeroImpute*. The overall classification accuracy by CMVE (See Fig. (f)) clearly shows higher classification accuracy for the range of missing values from 1-20% when values are imputed using CMVE as compared to other estimation algorithms. The reason for this better performance is that CMVE exploits all

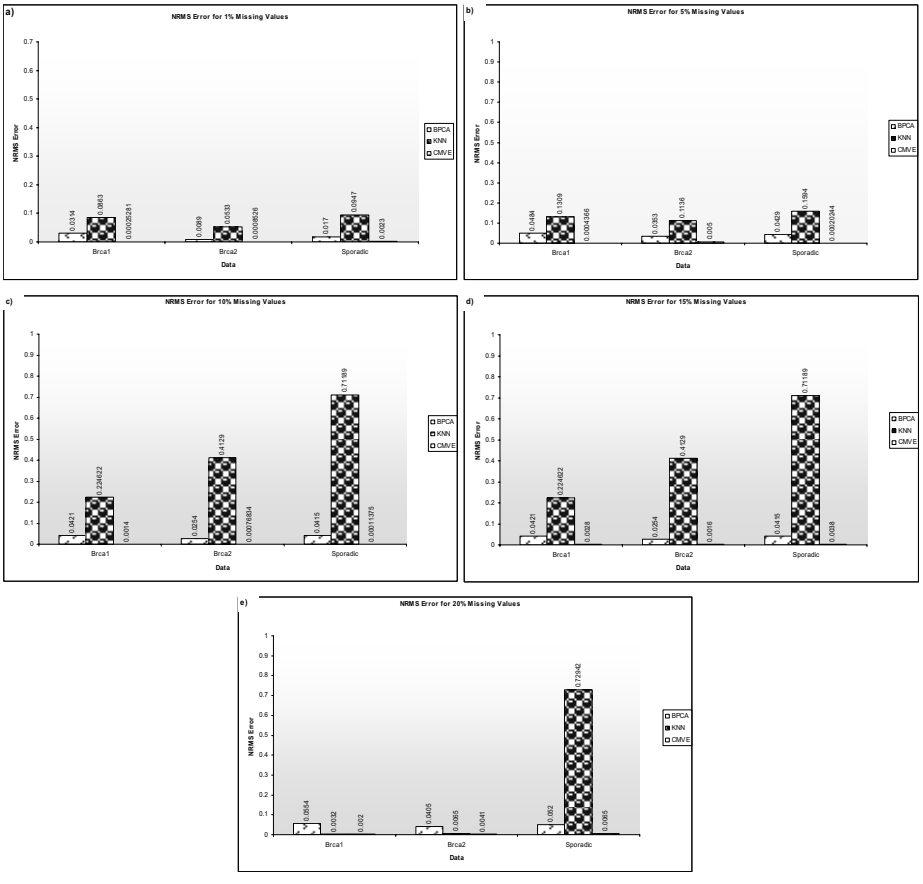

Fig. 1. NRMS Error for KNN, BPCA and CMVE for 1-20% Missing Values

types of correlation structure of the data as compared to KNN that only considers positive correlations, BPCA that only considers global correlation and *ZeroImpute* that doesn't consider any correlation.

Results in Fig. 2 (a-f) also draw attention to some interesting observation, RPLS followed by gene selection step, performed better when *ZeroImpute* was used as compared to KNN (See Fig. 2(f)). Because the data between classes was more separable and thus easier to classify, that is zero values actually improved separability. The other reason of poor performance of KNN is that if smaller k is used by KNN it increases the variance of the data leading to false selection of significant genes, however large value of k increases bias and leads to coarse estimates. In practice however, for the vast majority of datasets, zero imputation will not improve separability because for instance, if a particular gene has missing values, for both classes to be classified, *ZeroImpute* results in the same value, namely zero [5]. This means the gene has same value for both classes despite some genes being more significant than others. Also, BPCA performed better than KNN and *ZeroImpute* due to better estimation of missing values because of considering both positive and negative correlations [14].

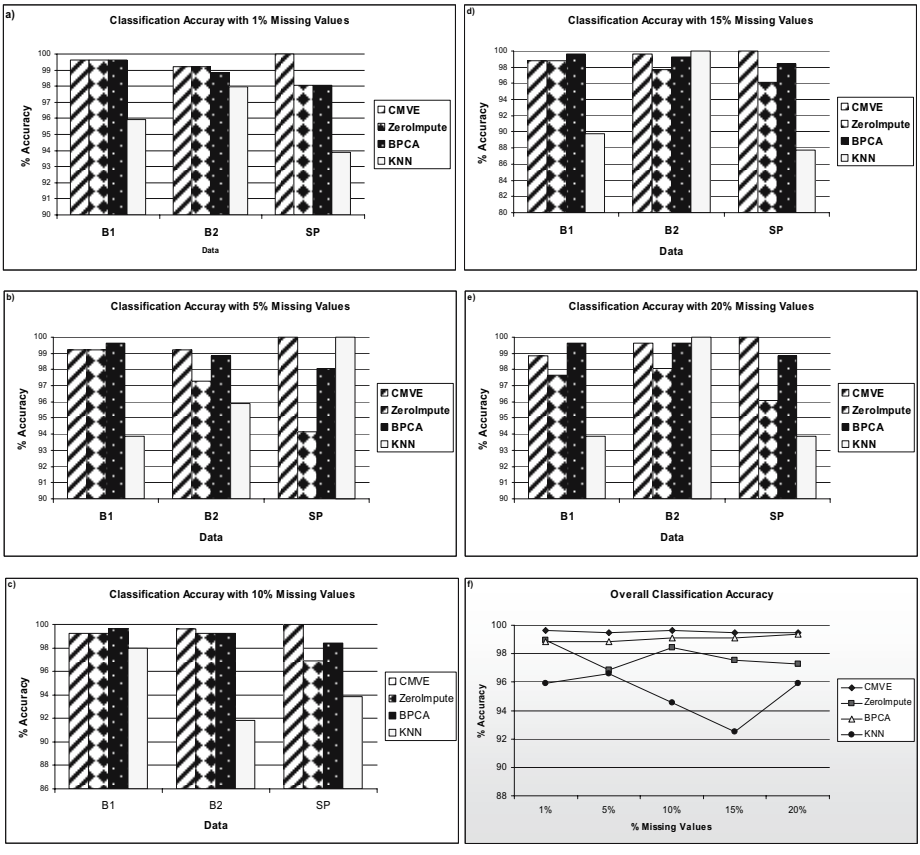

Fig. 2. Class Prediction Accuracy using CMVE, BPCA, KNN and *ZeroImpute* to estimate between 1% and 20% Missing Values

For this reason, it is always better to exploits all sort of correlation structure for estimation in the data.

4 Conclusions

This paper has presented a new *Collateral Missing Value Estimation* (CMVE) algorithm for accurate missing value estimation which leads to better gene selection and classification. CMVE has demonstrated superior imputation performance compared to the *Bayesian Principal Component Analysis* (BPCA), *K Nearest* Neighbour (KNN) algorithm and *ZeroImpute* methods, for estimating randomly missing values over the probability range from 0.01 to 0.2 in the BRCA1, BRCA2 and Sporadic genetic mutation samples present in ovarian cancer. Experimental results also reveal that CMVE consistently outperformed BPCA, KNN and *ZeroImpute* techniques in terms of their classification accuracies by exploiting all types of correlations between the data. The *Ridge Partial Least Squares* (RPLS) classifier was applied for the class prediction followed by the fusion of genes selection method, *Between Group to within Group*

Sum of Squares (BSS/WSS) and *Weighted Partial Least Square* (WPLS), and these afforded consistently improved classification performance for all experiments on ovarian cancer microarray data, when used in combination with CMVE. The results also corroborate the theoretical basis for the better performance of CMVE which means it can be successfully applied to any correlated data.

References

[1] M. S. B. Sehgal, I. Gondal, and L. Dooley, "Statistical Neural Networks and Support Vector Machine for the Classification of Genetic Mutations in Ovarian Cancer," *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)'04, USA.*, pp. 140-146, 2004.

[2] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. F. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," presented at Proc. Natl. Acad. Sci, USA, 2001.

[3] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A Bayesian Missing Value Estimation Method for Gene Expression Profile Data," *Bioinformatics*, vol. 19, pp. 2088-2096, 2003.

[4] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, pp. 77-78, 2002.

[5] M. S. B. Sehgal, I. Gondal, and L. Dooley, "K-Ranked Covariance Based Missing Values Estimation for Microarray Data Classification," *IEEE Hybrid Intelligent Systems (HIS)'04, Japan*, vol. 00, pp. 274-279, 2004.

[6] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect in the classifier accuracy," *Classification, Clustering and Data Mining Applications*, pp. 639-648, 2004.

[7] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman, "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics*, vol. 17, pp. 520-525, 2001.

[8] M. S. B. Sehgal, I. Gondal, and L. Dooley, "A Collateral Missing Value Estimation Algorithm for DNA Microarrays," *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), USA*, pp. 377-380, 2005.

[9] M. S. B. Sehgal, I. Gondal, and L. Dooley, "Support Vector Machine and Generalized Regression Neural Network Based Classification Fusion Models for Cancer Diagnosis," *IEEE Hybrid Intelligent Systems (HIS)'04, Japan*, pp. 49–54, 2004.

[10] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc, "Effective dimension reduction methods for tumor classification using gene expression data," *Bioinformatics*, vol. 19 no. 5, pp. 563-570, 2003.

[11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasen-beek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Down-ing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lan-der, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, pp. 286(5439):531-537, 1999.

[12] P. Broët, A. Lewin, S. Richardson, C. Dalmasso, and H. Magdelenat, "A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments," *Bioinformatics*, vol. 20, pp. 2562 - 2571, 2004.

[13] X. Liu, A. Krishnan, and A. Mondry, "An Entropy-based gene selection method for cancer classification using microarray data," *BMC Bioinformatics*, vol. 6:76, 2005.

[14] M. S. B. Sehgal, I. Gondal, and L. Dooley, "Collateral Missing Value Imputation: a new robust missing value estimation algorithm for microarray data," *Bioinformatics*, vol. 21(10), pp. 2417-2423, 2005.

[15] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent, "Gene-expression profiles in hereditary breast cancer," *N. Engl. J. Med*, pp. 22; 344(8):539-548, 2001.

[16] G. Fort and S. Lambert-Lacroix, "Classification using partial least squares with penalized logistic regression," *Bioinformatics*, vol. 21, pp. 1104-1111, 2005.

[17] M. Harvey and C. Arthur, "Fitting models to biological Data using linear and nonlinear regression," Oxford University Press, 2004.

[18] K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery, "Bayesian Model Averaging: development of an improved multi-class, gene selection and classification tool for microarray data," *Bioinformatics*, vol. 21 no.10, pp. 2394-2402, 2005.

[19] X. Zhou, X. Wang, and E. R. Dougherty, "Gene Selection Using Logistic Regressions Based on AIC, BIC and MDL Criteria," *New Mathematics and Natural Computation*, vol. 1, pp. 129-145, 2005.

[20] A. J. Amir, C. J. Yee, C. Sotiriou, K. R. Brantley, J. Boyd, and E. T. Liu, "Gene Expression Profiles of Brca1-Linked, Brca2-Linked, and Sporadic Ovarian Cancers," *Journal of the National Cancer Institute*, vol. 94 (13), 2002.