A COLLATERAL MISSING VALUE ESTIMATION ALGORITHM FOR DNA MICROARRAYS

Muhammad Shoaib B. Sehgal, Iqbal Gondal, Laurence Dooley
{Shoaib.Sehgal, Iqbal.Gondal, Laurence.Dooley}@infotech.monash.edu.au
GSCIT, Monash University, VIC 3842, Australia

ABSTRACT

Genetic microarray expression data often contains multiple missing values that can significantly affect the performance of statistical and machine learning algorithms. This paper presents an innovative missing value estimation technique, called *Collateral Missing Value Estimation* (CMVE) which has demonstrated superior estimation performance compared with the *K-Nearest Neighbour* (KNN) imputation algorithm, the *Least Square Impute* (LSImpute) and *Bayesian Principal Component Analysis* (BPCA) techniques. Experimental results confirm that CMVE provides an improvement of 89%, 12% and 10% for the BRCA1, BRCA2 and Sporadic ovarian cancer mutations respectively compared to the average error rate of KNN, LSImpute and BPCA imputation methods, over a range of randomly selected missing values. The underlying theory behind CMVE also means that it is not restricted to bioinformatics data, but can be successfully applied to any correlated data set.

1. INTRODUCTION

Microarrays are extensively used in the study of many biological processes varying from human tumours to yeast sporulation [1], with several mathematical, statistical processes and machine learning algorithms using such data for diagnosis and drug discovery. The most commonly used methods include clustering, classification and dimension reduction techniques such as *Principal Component Analysis* (PCA) and *Singular Value Decomposition* (SVD).

Despite wide usage, experimentally obtained microarray data frequently contains missing values with up to 90% of the genes affected by such missing values [2], which occur due to slide scratches, hybridization failures, image corruption or simply dust on slides [5]. Previous work has highlighted [3,4] that data dimension reduction techniques and machine learning algorithms including Support Vector Machines (SVM) and neural networks are affected by missing values in microarray data. The problem can be managed in many different ways from repeating the experiment which is not feasible for economic reasons, through to simply ignoring the samples containing missing values, though this often is inappropriate because usually there are a very limited numbers of samples available. The best solution is to estimate the missing values, but unfortunately most systems use zero impute (replace missing values by zero) or row average/median (replacement by the corresponding row average/median), neither of which exploit the correlation of data and result in high estimation errors [1]. Current research has demonstrated that if a correlation between the data is used then missing value prediction error can be reduced significantly [5]. Several methods including *K-Nearest Neighbour* (KNN) *Impute*, *Least Square Imputation* (LSImpute) [5] and *Bayesian Principal Component Analysis* (BPCA) [7] have been used. However, the prediction error produced by these methods still affects statistical and machine learning algorithms including class prediction, class discovery and gene identification algorithms. In these circumstances there is still a need to design a method which will provide minimal prediction error.

This paper presents a *Collimator Missing Value Estimation* (CMVE) technique which combines multiple value matrices for particular missing data. Different tests were conducted by randomly removing between 1% and 5% of values from the BRCA1, BRCA2 and Sporadic mutation microarray data (mutations present in ovarian cancer) [6] and then applying KNN, LSImpute, BPCA and CMVE to estimate the missing values. The *Normalized Imputation Root Mean Square* (NIRMS) error [2] was used to evaluate the performance of each estimation technique, with results demonstrating the superior performance of CMVE over the range of missing values and while it is not as critical as estimation accuracy, particularly when related to health care [1,5], the computational complexity order of CMVE is exactly the same as for the three algorithms mentioned above .

The rest of paper is organized as follows: Section 2 presents an overview of the three missing value estimation techniques used for comparative purposes, while the new CMVE algorithm is formally presented in Section 3. Section 4 analyzes the estimation performance of all four imputation methods, with some conclusions given in Section 5.

2. APPLIED MISSING VALUE ESTIMATION TECHNIQUES

2.1 K- Nearest Neighbour (KNN) Estimation

The KNN based method selects genes with expression values similar to the gene of interest to impute missing values [1]. In order to estimate the missing value Y_{IJ}, of gene I and experiment J, k genes are selected whose expression vectors are similar to genetic expression of I in samples other than J. The similarity measure between two expression vectors Y_1 and Y_2 is defined by the reciprocal of the Euclidian distance over the observed components in experiment J.

$$\psi = 1/\| Y_1 - Y_2 \| \tag{1}$$

The missing value is then estimated as the weighted average of the corresponding entries in the selected k expression vectors:-

$$\hat{Y}_{IJ} = \sum_{i=1}^{k} W_i . X_i \tag{2}$$

$$W_i = \frac{1}{\psi_i \times \Delta} \tag{3}$$

where $\Delta = \sum_{i=1}^{k} \psi_i$, ψ is the Euclidean distance and X is the input matrix containing gene expressions. (2) and (3) show the contribution of each gene is weighted by the similarity of its expression to gene I.

KNN based imputation method has no theoretical criteria for selecting the best k-values which are empirically determined. Also, the Euclidean distance measure is sensitive to outliers, who may be present in microarray data, though our research showed that log-transformation of the data significantly reduced the effects of outliers on gene similarity determination. The choice of a small k degraded the performance of the classifier as the imputation process overemphasized a few dominant genes in estimating the missing values. Conversely, a large neighbourhood would include genes that may be significantly different from those containing missing values, so degrading the estimation process and commensurately the classifier's performance. Our empirical results demonstrated that for small datasets, $k=10$ was most effective confirming the observation in [4].

2.2 Least Square Impute

Least Square Impute (LSImpute) is a regression based missing data estimation method which exploits the correlation between genes. To estimate the missing value $Y_{IJ,}$ of gene I from gene expression matrix X containing non-missing values for gene I and experiment J, firstly the k most correlated genes are selected whose expression vectors are similar to gene I from X in experiments other than J. The regression method is then used to estimate value Φ_1 for Y_{IJ} as

$$\Phi_1 = \alpha + \beta X + \xi \tag{4}$$

where ξ is the error term for which the variance is minimized when least squares (LS) estimating the model (parameters α and β). In single regression, the estimate of α and β gives

$$\alpha = \bar{y} - \beta \bar{X} \text{ and } \beta = \frac{\Im_{xy}}{\Im_{xx}}$$

where $\Im_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y})$ is the empirical covariance between X and Y, $\Im_{xx} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$ is the empirical variance of X and n the number of samples. Here $\bar{X}$ and $\bar{Y}$ are the means over $X_1,...,X_n$ and $Y_1,...,Y_n$. Thus the LS estimate of a variable Y given a variable X can be written as $\hat{Y} = \bar{Y} - \frac{\Im_{xy}}{\Im_{xx}}(X - \bar{X})^2$.

2.3 Bayesian Principal Component Analysis based Estimation

Bayesian Principal Component Analysis (BPCA) estimates missing values Y^{miss} of data matrix Y using Y^{obs}. The probabilistic PCA (PPCA) is calculated using the Bayes theorem and the Bayesian estimation calculates posterior distribution of θ and X using:-

$$p(\theta, X \mid Y) \ \alpha \ p(Y, X \mid \theta) p(\theta) \tag{5}$$

where $p(\theta)$ is called the *prior distribution* which donates a priori preference for parameter θ and X is the input matrix containing gene expression samples.

The missing values are estimated using a recursive algorithm which works as follows: Bayesian estimation (BE) is executed for both model parameter θ and Y^{miss} like expectation maximization repetitive algorithm and calculates the posterior distributions for θ and Y^{miss}, $q(\theta)$ and $q(Y^{miss})$, by a repetitive algorithm as in [7]. Finally, missing values in gene expression matrix are imputed using

$$\hat{Y} = \int Y^{miss} q(Y^{miss}) dY^{miss} \tag{6}$$

$$q(Y^{miss}) = p(Y^{miss} \mid Y^{obs}, \theta_{true}) \tag{7}$$

where θ true is the posterior of the missing value.

3. COLLATERAL MISSING VALUE ESTIMATION (CMVE) ALGORITHM

The CMVE algorithm is based on a concept of multiple parallel estimations of missing values. For example, if value Y_{IJ}, of gene I and sample J is missing and CMVE estimates multiple values for it, then based on these values the final value χ for Y_{IJ} is estimated. The complete CMVE algorithm is shown in Fig. 1. Firstly the diagonal covariance of I is computed together with the other gene expressions, where N is the number of genes and I the gene number with missing value for sample J. Rows are then sorted according to their covariance, with the first *k-ranked covariate* genes being selected. The reason for using a covariance instead of a distance function, as was used by KNN, is explained by:

Lemma 1: Distance functions only consider positive correlations.

Proof: If there are two sets α and β which are inversely proportional to each other, then the distance d between α and β will be larger in those sets which are directly proportional to each other. Several distance functions are used for KNN, the most common being Gaussian which is given by:-

$$d = \|\alpha - \beta\| \tag{8}$$

which always gives a higher value of d when α is inversely proportional to β.

Lemma 2: The CMVE algorithm considers both positive and negative correlation values.

Proof: Assume two sets α and β that are inversely proportional, so $cov < 0 \forall \alpha, \beta$ where

$$\text{cov} = \frac{1}{(n-1)}\sum_{i=1}^{k}(\alpha_i - \bar{\alpha})(\beta_i - \bar{\beta}) \tag{9}$$

From (9), it is clear that if a high correlation exists between the gene values (either directly proportional and positively correlated values or inversely proportional and negatively correlated values) a higher absolute *cov* value will always be generated.

Let Φ_1 be the estimate of Y_{IJ} in (4) (Step 4a) using the linear regression method in Section 2.2, while Step 4b estimates two other sets of missing values Φ_2 and Φ_3. Φ_2 is estimated using:-

$$\Phi_2 = \sum_{i=1}^{k} \phi + \eta - \sum_{i=1}^{k} \xi^2 \quad (10)$$

Similarly value of Φ_3 is computed using:-

$$\Phi_3 = \frac{\sum_{i=1}^{k} (\phi^T \times I)}{k} + \eta \quad (11)$$

η and ϕ in (10) and (11) are obtained from the Non Negative Least Square (NNLS) method [4]. The aim is to find a linear combination of models, that best fit R_k and I. The objective function in NNLS is used to minimize the prediction error ξ_0 as:

$$\phi, \eta = \min(\xi_0) \quad (12)$$

Linear programming is used to find coefficients ϕ which have minimum prediction error and residual η. The value of ξ_0 in (12) is calculated using:-

$$\xi_0 = \max(SV(R_k.\phi - I)) \quad (13)$$

where SV are the singular values of the difference vector between product R_k and prediction coefficients ϕ with the gene expression row I containing missing values. The tolerance used by the linear programming method to compute vector ϕ is:-

$$Tol = k \times N \times \max(SV(R_k)) \times C \quad (14)$$

where k = number of predictor genes R_k and C is the number of predictor gene samples. Finally, value χ for Y_{IJ} is computed using:-

$$\chi = \alpha.\Phi_1 + \beta.\Phi_2 + \gamma.\Phi_3 \quad (15)$$

where the values of α, β and γ are set to 0.33 to obtain the average of Φ_1, Φ_2 and Φ_3.

Pre Condition: Gene expression matrix *G(R,N)* with *R* number of genes, *N* samples, *I* missing values, *index*=1
Post Condition: *G* without any missing values.

Algorithm:
1- Compute absolute covariance *CoV* using (9)
2- Rank genes (rows) based on *CoV*
3- Select the *k* most effective rows R_k
4- Use values of R_k to
   a. Estimate value Φ_1 using (4)
   b. Compute Φ_2 and Φ_3 using (10) and (11)
5- Compute missing value of *I[index]* using (15)
6- Impute estimated value χ in (15) and use in future predictions
7- Increment *index* and Repeat Steps 1–6 until all missing values of *G* are estimated

Fig.1: Collateral missing value estimation algorithm

Two further features of CMVE are now discussed which underscore the superior performance of this algorithm.

Lemma 4: Prediction error probability of CMVE will always be less than BPCA, LSImpute and KNN.

Proof: The prediction error probability is directly proportional to the number of missing values *M* [8] for correlated data. Assume P_1 and P_2 are the prediction error probabilities of the comparative methods (BPCA, LSImpute and KNN) and CMVE respectively, where ε_1 and ε_2 are the actual prediction errors such that:-

$$P_1 = \sum_{i=0}^{M} P(\varepsilon_1) p(M) \quad (16)$$

$$P_2 = \sum_{i=0}^{M} P(\varepsilon_2) P(i) \quad (17)$$

P_1 is the summation of the product of prediction error probabilities and probability of missing values since the comparative methods do not consider estimated values in future missing value predictions and such algorithms only consider *M* missing values for each prediction. In contrast, CMVE uses estimated values for future prediction of missing values so each estimate increases the predictor genes to be considered and decreasing the prediction probabilities in (17) so:

$$P_1 < P_2 \text{ such that } P_2 \to 0 \text{ when } i \to 0 \quad \because P(i) = 0 \text{ for } i = 0 \quad (18)$$

Lemma 5: CMVE always provides a better estimate of missing values in the case of transitive gene dependency (Gene A→B→C) than BPCA, LSImpute and KNN.

***Proof*:** Assume that gene G_{a1} is correlated with S_1 such that:-

$$G_{a1} \to S_1 \text{ such that } S_1 = \{G_{b1}, G_{b2} ... G_{bn}\} \quad (19)$$

Similarly gene G_{b1} is correlated with S_2 as:-

$$G_{b1} \to S_2 \text{ such that } S_2 = \{G_{c1}, G_{c2} ... G_{cn}\} \quad (20)$$

If the values of both G_{a1} and G_{b1} are missing then G_{b1} can be predicted using set S_2 and then subsequently used to predict G_{a1} more accurately using S_1 by including G_{b1} rather than ignoring it. Unlike CMVE, all the aforementioned techniques do not consider estimated values in predicting future missing values.

For completeness the computation complexity order of all missing values considered is undertaken as follows:

Lemma 6: Computational complexity order of CMVE is exactly the same as for KNN, LSImpute and BPCA algorithms.

Proof: The critical operation for CMVE, KNN, LSImpute and BPCA is to search for the most correlated values. So, CMVE has same complexity order as KNN, LSImpute and BPCA. The added computation overhead for multiple imputations by CMVE is negligible as compared to searching for the most correlated genes.

5. DISCUSSION OF RESULTS

To test different imputation algorithms, microarray data by Amir et al [6] was used in all experiments. 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. The missing value estimation techniques were tested by randomly removing data values and then computing the estimation error. For test purposes, between 1% and 5% of values were removed from each dataset samples and the NIRMS errors ξ computed as:

$$\xi = \frac{RMS(M - M_{est})}{RMS(M)} \quad (21)$$

where M is the original data matrix and M_{est} is the estimated matrix using KNN, LSImpute, BPCA and CMVE. The motivation for using this metric for error estimation is that ξ =1 for zero imputation [2].

Fig. 2 to 6 show the NIRMS error for a range of imputes, randomly missing values and results prove that CMVE out-performed all other techniques not only for a small number of missing values, but also higher values. For example, the average improvement in performance was 94%, 95% and 93 % for 4% missing values and 93%, 94% and 92% for 5% missing values for the three genetic datasets respectively. This underscores the capability of the CMVE algorithm to more effectively estimate higher missing values for the reasons detailed in Lemma 4 and 5.

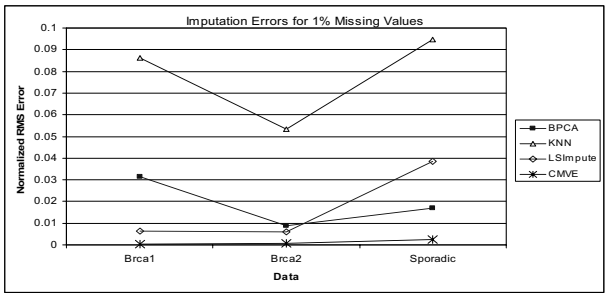

Fig. 2: Missing value imputation error for 1% missing values

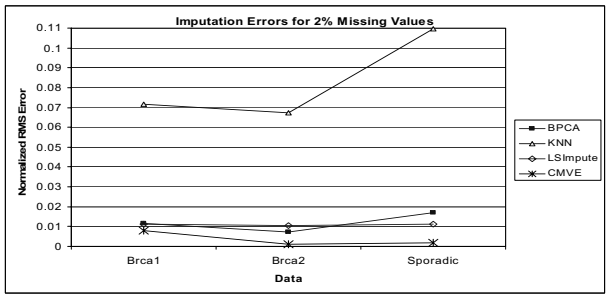

Fig. 3: Missing value imputation error for 2% missing values

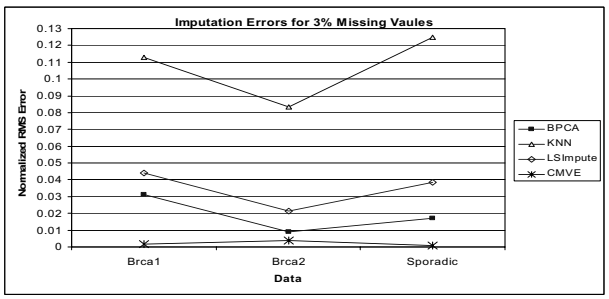

Fig. 4: Missing value imputation error for 3% missing values

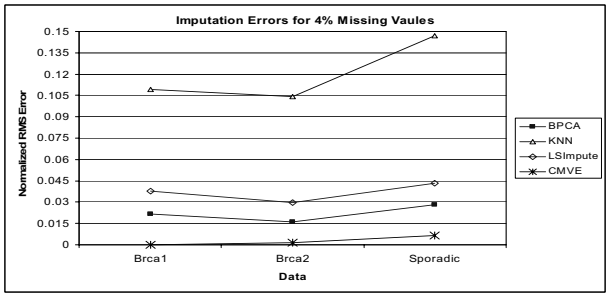

Fig. 5: Missing value imputation error for 4% missing values

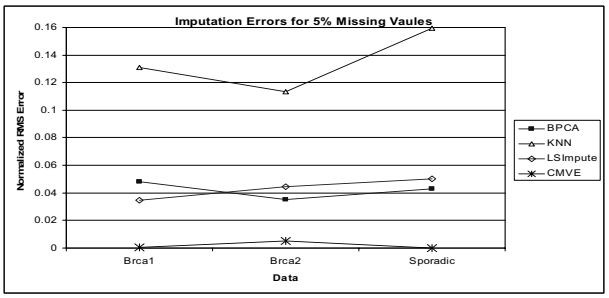

Fig. 6: Missing value imputation error for 5% missing values

6. CONCLUSIONS

This paper has presented a novel *Collateral Missing Value Estimation* (CMVE) algorithm, whose performance has been proven to be superior in terms of error rates, to other commonly used techniques including KNN, LSImpute and BPCA. Results confirmed that for randomly missing values between 1% and 5% on ovarian cancer microarray data, the overall performance improvement was on average 89%, 12% and 10% respectively for BRCA1, BRCA2 and Sporadic mutation data. CMVE also consistently demonstrated better performance for higher numbers of missing vales, with no overall increase in the order of computational complexity. While analysis has focused upon ovarian cancer microarray data, the algorithm's performance in minimizing estimation errors means it can be applied effectively to other datasets comprising correlated values.

7. REFERENCES

[1] 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.

[2] M. Ouyang, W.J. Welsh, P. Georgopoulos, "Gaussian Mixture Clustering and Imputation of Microarray Data", *Bioinformatics*, 2004.

[3] M.B.S. Shoaib, I. Gondal, L.Dooley, (2004), "Statistical Neural Networks and Support Vector Machine for The Classification of Genetic Mutations in Ovarian Cancer", to be published in Proc. IEEE CIBCB 04.

[4] 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.

[5] Bo TH, B. Dysvik, I. Jonassen, "Lsimpute: Accurate Estimation of Missing Values in Microarray Data with Least Squares Methods", *Nucleic Acids Res.,* pp. 32(3):e34, 2004.

[6] A.J. Amir , C. J. Yee, C. Sotiriou, K. R. Brantley, J. Boyd, E. T. Liu, "Gene Expression Profiles of Brca1-Linked, Brca2-Linked, and Sporadic Ovarian Cancers" *Journal of the National Cancer Institute*, vol. 94 (13), July 3, 2002.

[7] S. Oba, M.A. Sato, I. Takemasa, M. Monden, K. Matsubara, S. Ishii, "A Bayesian Missing Value Estimation Method for Gene Expression Profile Data", *Bioinformatics*, vol.19 (16), pp.2088-2096, 2001.

[8] A. McLean, "The Predictive Approach to Teaching Statistics", *Journal of Statistics Education,* v.8, n.3, 2000.