ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

MISSING VALUE ESTIMATION OF EPISTATIC MINIARRAY PROFILING DATA BY KERNEL PCA REGRESSION ENSEMBLE APPROACH

Malathi.M[1], Dr. Antony Selvadoss Thanamani[2]

Research Scholar, Dr. Mahalingam Centre for Research and Development, NGM College, Pollachi, India[1]

Associate Professor and Head, Dr. Mahalingam Centre for Research and Development, NGM College, Pollachi, India [2]

Abstract: Missing data imputation is a key issue in learning from incomplete data. Various techniques have been developed with great success on dealing with missing values in data sets with heterogeneous attributes (their independent attributes are of different types) referred to as imputing mixed-attribute data sets. Epistatic miniarray profiling (E-MAP) is a powerful tool for analyzing gene functions and their biological relevance. However, E-MAP data suffers from large proportion of missing values, which often results in misleading and biased analysis results. This paper studies a new setting of missing data imputation, a novel ensemble approach EMDI based on the high-level diversity to impute missing values that consists of two global (kernel principal component analysis regression) and four local base estimators. The performance of the proposed KPCAR impute algorithm is compared with state-of-the-art linear regression methods, i.e., Bayesian principal component analysis imputation (BPCA). The KPCAR impute outperforms the L-Simpute when the missing percentage increases. The performance of the KPCA impute is similar to that of the BPCA imputation. Therefore, it is an effective and promising algorithm in estimating missing values for DNA microarray profiles.

Keywords: Epistatic miniarray profiling, missing value estimation, Matrix completion, KPCA.

I. INTRODUCTION

Epistatic Miniarray profiling (E-MAP) [1,2] is a tool for generating and analysing comprehensive genetic-interaction maps systematically, and it is a variant of synthetic genetic array [3], which is a high throughput technique for functional analysis. It measures the strength of pair-wise genetic interactions quantitatively using continuous real values, i.e., negative value indicates two genes working in compensatory pathways and positive value exclusively working in the same pathway [4]. It is usually represented in the form of a symmetric matrix but often with a large number of missing values, some of which lose as much as 30% data. Although it has been generally acknowledged that valuable insights into the underlying biological process can be obtained via systematically analysing E-MAP scores using methods like clustering and matrix factorization, these algorithms significantly suffer from the missing information and often produce biased outputs [5].

Hence, in order to improve the subsequent data analysis quality and provide more direct knowledge of the tested genes, it is highly desired to accurately impute the missing values in E-MAP.

When faced with incomplete data in E-MAP, data analysts typically have the following options before applying the following popular data analysis tools: (1) discard the genes that contain missing values; (2) replace missing values with some constants (e.g. zero); and (3) estimate missing values using some sophisticated imputation approaches. E-MAP dataset contains very large percentage of missing values and almost every gene has more or less missing values, in the meanwhile, it is also a symmetric matrix, which brings about the result that each missing value requires removal of two genes. So if we take option (1), more than 90% gene data in the E-MAP matrix could be removed. Hence simply discarding the genes with missing data is not applicable. It is clearly unreasonable to impute all missing values with the same constant because of gene specificity. For that matter, option (2) is also not an optimal approach. To analyse E-MAP score data more accurately, option (3) is a better choice and can make recovered missing values more close to real values by exploiting the information buried in the non-missing part. It has been proved that sophisticated imputation method is useful in reducing missing value's impact and making the subsequent analysis to be as informative as possible [5].

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

Missing data imputation aims at providing estimations for missing values by reasoning from observed data [6]. Because missing values can result in bias that impacts on the quality of learned patterns or/and the performance of classifications, missing data imputation has been a key issue in learning from incomplete data. Various techniques have been developed with great successes on dealing with missing values in data sets with homogeneous attributes (their independent attributes are all either continuous or discrete). However, these imputation algorithms cannot be applied to many real data sets, such as equipment maintenance databases, industrial data sets, and gene databases, because these data sets are often with both continuous and discrete independent attributes [7]. These heterogeneous data sets are referred to as mixed-attribute data sets and their independent attributes are called as mixed independent attributes in this research.

Imputing mixed-attribute data sets can be taken as a new problem in missing data imputation because there is no estimator designed for imputing missing data in mixed attribute data sets. The challenging issues include, such as how measuring the relationship between instances (transactions) in a mixed-attribute data set, and how to construct hybrid estimators using the observed data in the data set. To address the issue, this research proposes a non parametric iterative imputation method based on a mixture kernel for estimating missing values in mixed-attribute data sets. It first constructs a kernel estimator to infer the probability density for independent attributes in a mixed-attribute data set. And then, a mixture of kernel functions (a linear combination of two single kernel functions, called mixture kernel) is designed for the estimator in which the mixture kernel is used to replace the single kernel function in traditional kernel estimators.

II. RELATED WORK

Methods for dealing with missing values can be classified into three categories by following the idea from [8]: 1) case deletion, 2) learning without handling of missing values, and 3) missing value imputation. The case deletion is to simply omit those cases with missing values and only to use the remaining instances to finish the learning assignments [9]. The second approach is to learn without handling of missing data, such as Bayesian Networks method [10], Artificial Neural Networks method 15, the methods in [11]. Different from the former two, missing data imputation method advocates filling in missing values before a learning application. Missing data imputation is a procedure that replaces the missing values with some plausible values, such as [12]. While the imputation method is regarded as a more popular strategy, a new research direction, the par imputation strategy, has recently been proposed in [13]. It advocates that a missing datum is imputed if and only if there are some complete instances in a small neighbourhood of the missing datum, otherwise, it should not be imputed.

A. Matrix completion (MC)

The natural assumption that the underlying matrix is low-rank; we can recover missing value of matrix from a small number of observed entries, which is known as "compressed sensing" or "matrix completion" that recently attracted substantial attention. Suppose that we have an unknown matrix $M \in R^{m \times n}$ of rank at most r, given an observed subset $E \subseteq [m] \times [n]$, the low-rank matrix completion problem is to find a matrix $X \in R^{m \times n}$ with the minimum rank. It can be formulated as solving the following optimization problem:
Minimize rank(X); subject to $X_{i,j} = \mathbf{M}_{i,j}; (i,j) \in E$
Where rank(X) is the rank of matrix X.

B. Bayesian principal component analysis (BPCA)

BPCA is an estimation method based on probabilistic model, which exploits a variety of Bayesian algorithms, such as principle component regression, Bayesian estimation, and Expectation Maximization (EM), to iteratively maximize posterior distribution of model parameters and missing values until convergence. The algorithm automatically obtains the most relevant principal axes which are used for regression and shrinks other redundant axes toward 0. The whole process is as follows: it initially imputes missing values using row average, then the posterior distributions of the parameter and missing values are maximized alternately. At last missing values are estimated using the expectation of missing value posterior distribution.

III. OUR CONTRIBUTION

This paper proposes a better accuracy for missing value estimation of Epistatic miniarray profiling data by novel ensemble approach using kernel principal component Analysis Regression. A unique feature of this method is that the computed KNN algorithm, Local Lease square, Lease squares impute and missing value imputation ensemble such as those in five E-MAP datasets.

The degree (the number of other genes that one gene has known interacting with) distribution of the 5 different E-MAP datasets, and we can see that the degree distribution of Chromosome dataset is more uniform compared to other datasets. For dataset with uniform degree distribution and large percentage of missing values, estimation using linear regression of *K* coherent genes is not effective. When the degree distribution is not uniform, LLS performs better.

A. *Principal Component Analysis*

PCA represents a matrix of process variables as the product of two matrices, one containing the transformed variables (scores), and the other containing the new axes of rotation (loadings or projection directions). Given a n x r

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

matrix of measured process variables = X' + εx, where X' is the matrix of underlying noise-free data, εx is the additive noise matrix, r is the number of variables, and n is the number of observations, PCA decomposes the matrix X as,

$$X = Z a^T \qquad (1)$$

Where **Z** is a n x r matrix of the principal components or the principal component scores, and **a**is an orthogonal r ´ r matrix of the loadings or projection directions. This transformation diagonalizes the data covariance matrix as

$$X^T X = aDa^T \qquad (2)$$

Where **D** is a diagonal matrix containing the eigenvalues of the data covariance matrix. Substituting Equation (1) into Equation (2) gives

$$X^T X = (Za^T)^T Za^T = aZ^T Za^T = aDa^T \qquad (3)$$

This indicates that the principal components are uncorrelated variables with variances equal to the eigenvalues of the data covariance matrix.

B. K-nearest Neighbours (KNN) Algorithm

KNN is an easy and widely used imputation method. It first finds K most similar genes of the query gene with missing value and then estimates the missing value as the average of the K most similar genes (uKNN) or weighted by the inverse of their distance (wKNN) to the query gene. In method uKNN, for each missing interaction (*i* and *j*), *K* most similar genes of both gene i and gene j are found, and then missing value is estimated as the average of these values. wKNN is a variant of uKNN. It considers the effect of similarity extent between genes; namely, more similar gene makes a greater contribution to the imputation. In this study, like weighting method described in [14], given gene i and its neighbor k, the weight is computed as follows:

$$w(i,k) = \left(\frac{\delta^2}{1-\delta^2+\varepsilon}\right)^2 \qquad (4)$$

Where $\varepsilon=10^{-6}$ and δ is the correlation coefficient between genes *i* and *k*, which ensures that the closer the neighbors.

C. Local least square

LLS is a method based on least squares, and it selects *K* coherent genes, instead of weighting or averaging *K*most similar genes, multiple regression using *K* neighbors is performed based on the following formula:

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k \qquad (5)$$

Where x_i is the y^{th} nearest neighbor of query gene, αi is the regression coefficient and it is determined by pseudo inverse of the *K* neighbor's expression matrix, whose missing values are initially assigned with row average.

D. Least squares impute

LSimpute is a combination of gene-based and array-based imputation methods, where gene-based estimate is imputed as weighted average of the predicted values from regression model which is built over K most nearest genes and it is based on gene correlation. Array-based estimate is determined by multiple regressions on the arrays like LLS and it is based on array correlation. After obtaining the estimated values of gene-based yg and array-based ya, the combined estimate is calculated as $y=\beta y_g+ (1-\beta)y_a$, where mix coefficient β is determined by minimizing the sum of squared errors for artificial missing values.

E. Kernel principal component analysis Regression (KPCAR)

Both BPCA and LSimpute method assume the linear relations between genes. However, it is almost impossible to know exactly the relations to be linear. Under such circumstances KPCA regression techniques is explored, since it captures both linear and non-linear relations.

Kernel principal component analysis regression (kernel PCA) is a generalization of standard PCA. It effectively exploits the "kernel trick" to find the features of observation data. In kernel PCA regression, the regressors are not the observation data in the input space, but a nonlinear mapping of the observation data into the feature space. The regressors are thus called the features of the observation Data. To avoid non- linear mapping, a kernel function has been defined in the input space. A kernel matrix can be generated, of which each element is defined by the kernel function. The standard PCA has been performed on the kernel matrix such that the principal components of the features are first determined. They can be used as the regressors.

The general KPCA regression model is written in the following manner

$$yi = f(X_i') + e_i = f(x_{i1}, x_{i2}, \ldots, x_{ik}) + e_i, i = 1 \ldots n \qquad (6)$$

Where f(·) is a smooth and continuous but unspecified function and ei is the normally distributed error with mean 0 and variance σ^2. Moreover, the objective of KPCA regression is to estimate the regression function f(·) directly rather than to estimate parameters.

The estimation ability of KNN and Lease square impute methods depends on important model parameters, such as the k-value in KNN impute and the number of eigenvectors in Least square impute. This is a global method consisting of three components. First, the principal component regression, which is basically a low rank approximation of the data set, is performed. Second, the Kernel estimation, which assumes that the residual error and the projection of each gene on principal components behave as normal independent random variables with unknown parameters, is carried out. Third, regression estimation follows by iterations based on the expectation-maximization (EM) of the unknown Bayesian parameters.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

The KPCA take advantage of the particular properties of the data. In the method, they are compared at different percentages of missing data in terms of the similarity between the original and imputed data.

IV. IMPLEMENTATION

The missing value estimation algorithm has also been implemented in Matlab: Version 2010a. The implementation steps are:

Step1: Data formats: The information buried in the non-missing part of the dataset to estimate missing values have been developed and successfully applied. They generally fall into trained (supervised) and non-trained (unsupervised) groups. The trained methods usually use machine learning algorithms, which construct a predictive model to estimate absent values using data from the remaining attributes. The non-trained methods do not use target values, such as imputing with mean values, or based on parameter estimation statistical methods by maximum likelihood (variants of the Expectation–Maximization algorithms).The trained imputation methods and it also can be further categorized into two subgroups: (1) local-based imputation methods, and (2) global-based imputation methods.
Step2: Pre-processing: In this step, the pre-processing stage checks, PCA requires mean centring, because it is based on the calculation of the covariance matrix. Thus, the mean must be subtracted before estimating missing values and added again afterwards. This is done automatically by the methods presented here.

Step3: KNN: The KNN impute finds K other genes with the closest similar gene profiles using Pearson's correlation or Euclidean distance, then the missing value is estimated by the weighted or unweighted average of values from the selected *K* other closest genes.

Step4: KPCA: KPCA initially sets missing values as row mean, and then estimation methods based on probabilistic model, such as principle component analysis, regression estimation, and expectation maximization, are employed until convergence is reached for refinement.

Step5: Evaluation: The Evaluation returns correlation coefficient and nrmse for real values and its estimated values, the imputed matrix using missing value imputation algorithm.

V. RESULTS

To evaluate the performance of imputation methods, we artificially introduce additional 1% missing values on all the datasets. Among the 1% missing values, we hide the interaction value of two genes (A and B) in E-MAP score matrix and its symmetrical interaction value (B and A) because of E-MAP's symmetry property, which conforms to the real situation. To reduce bias in one test, this process is repeated 2 the performance on a different additional missing number rate 0.5%–6% in five E-MAP datasets.

The proposed method, apply two measures for evaluation, the first one is Pearson Correlation Coefficient (CC) between the predicted and actual interactions and given by:

$$CC = \frac{\sum_{i=1}^{L}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^{L}(x_i - \bar{x})^2][\sum_{i=1}^{L}(y_i - \bar{y})^2]}} \qquad (7)$$

Where x_i and y_i are the known and predicted values respectively, x and $y$are the corresponding means of known and predicted values, L is the number of hidden values for evaluation.

The second one is Normalized Root Mean Squared Error (NRMSE) measure and given by:

$$NRMSE = \sqrt{\frac{mean\,[(Q - P)^2]}{var\,(P)}} \qquad (8)$$

Where Q and P are the known and predicted value vectors. The higher the *CC*, the more accurate the imputation is. Contrary to *CC*, the lower the *NRMSE*, the more accurate the imputation is.

TABLE: 1 PERFORMANCE MEASURED BY PEARSON CORRELATION COEFFICIENT (CC) OF CHOOSING DIFFERENT REFERENCE ALGORITHMS IN THE ENSEMBLE ESTIMATOR

Algorithm	ESP	Pombe	Chromosome
KNN	0.708	0.739	0.649
LSimpute	0.700	0.730	0.640
BPCA	0.704	0.736	0.647
KPCAR	**0.765**	**0.810**	**0.785**

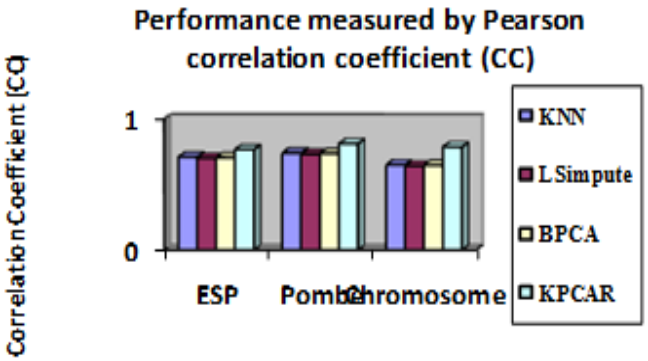

Fig 1: The Performance measured using Pearson CC

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

TABLE: 2 PERFORMANCE MEASURED BY NRMSE OF CHOOSING DIFFERENT REFERENCE ALGORITHMS IN THE ENSEMBLE ESTIMATOR

Algorithm	ESP	Pombe	Chromosome
KNN	0.713	0.679	0.766
LSimpute	0.714	0.685	0.773
BPCA	0.715	0.736	0.767
KPCAR	**0.705**	**0.660**	**0.710**

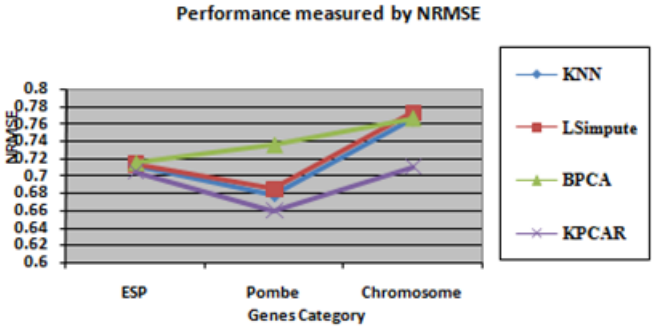

Fig 2: Graph showing the performances measured by NRMSE

VI. CONCLUSION

In this paper ensemble method is emphasized is regard with imputation of missing values with KPCA ensemble diversity for E-MAP missing value imputation. This method proved to have high performance and robustness compared to other traditional approaches. The kernel-based iterative regression estimators are proposed against the case that data sets have both continuous and discrete independent attributes. It utilizes all available observed information, including observed information in incomplete instances (with missing values), to impute missing values, whereas existing imputation methods use only the observed information in complete instances (without missing values).Experimental results indicate that our strategy is promising for solving the trouble of how to choose an optimal imputation for their specific data. For modelling the real situation of E-MAP dataset, artificially introduce additional missing values via hiding the value and its symmetrical value because of symmetry in E-MAP score matrix.

VII. FUTURE WORK

In future we have planned to find the solutions for biological information such as Gene Ontology (GO) Annotation or protein information will be integrated into the Collaborative Filtering Based on Rough-Set Theory approach to impute missing values in microarray datasets.

REFERENCES

[1] S.R. Collins, et al., Functional dissection of protein complexes involved in yeast chromosomebiology using a genetic interaction map, Nature 446 (2007) 806–810.
[2] M. Schuldiner, et al., Exploration of the function and organization of the yeast early secretory pathway through an epistaticminiarray profile, Cell 123 (2005) 507–519.
[3] A.H. Tong, et al., Global mapping of the yeast genetic interaction network, Science 303 (2004) 808–813.
[4] D. Fiedler, et al., Functional organization of the S. cerevisiae phosphorylation network, Cell 136 (2009) 952–963.
[5] T. Johannes, E. Laura, N. Olli, A. Tero, Missing value imputation improves clustering and interpretation of gene expressionmicroarray data, BMC Bioinform. 9 (2008) 202.
[6] G. Batista and M. Monard, "An Analysis of Four Missing Data Treatment Methods for Supervised Learning," Applied Artificial Intelligence, vol. 17, pp. 519-533, 2003.
[7] K. Lakshminarayan et al., "Imputation of Missing Data in Industrial Databases," AppliedIntelligence, vol. 11, pp. 259-275, 1999.
[8] K. Cios and L. Kurgan, "Knowledge Discovery in Advanced Information Systems," Trends in DataMining and Knowledge Discovery, N. Pal, L. Jain, and N. Teoderesku, eds., Springer, 2002
[9] K. Lakshminarayan et al., "Imputation of Missing Data in Industrial Databases," Applied Intelligence, vol. 11, pp. 259-275, 1999.
[10] R. Marco, "Learning Bayesian Networks from Incomplete Databases," Technical Report kmi-97-6, Knowledge Media Inst., The Open Univ., 1997.
[11] U. Dick et al., "Learning from Incomplete Data with Infinite Imputation," Proc. Int'l Conf.Machine Learning (ICML '08), pp. 232-239, 2008.
[12] Q.H. Wang and R. Rao, "Empirical Likelihood-Based Inference under Imputation for MissingResponse Data," Annals of Statistics, vol. 30, pp. 896-924, 2002.
[13] S.C. Zhang, "Parimputation: From Imputation and Null-Imputation to Partially Imputation," IEEE Intelligent Informatics Bull., vol. 9, no. 1, pp. 32-38, Nov. 2008.
[14] T. Bo, B. Dysvik, I. Jonassen, LSimpute: accurate estimation of missing values in microarraydata with least squares methods, Nucleic Acids Res. 32 (2004) e34.

BIOGRAPHIES

M.Malathi received her B.C.A and Master of Computer Applications from NGM College, Pollachi, Coimbatore, India. Presently she is working as an Assistant Professor in the Department of ComputerScience in NGM College (Autonomous), Pollachi.Her area of interest includes data Mining. Now she ispursuing her M.Phil Computer Science in NGM College ,Bharathiar University. Her area of Interests Data mining, Knowledge Engineering and Image Processing.

Dr. Antony SelvadossThanamani is presently working as Professor and Head, Dept of Computer Science, NGM College, Coimbatore, India (affiliated to Bharathiar University, Coimbatore). He has published more than 100 papers in international/ national journals and conferences. He has authored many books on recent trends in Information Technology. His areas of interest include E-Learning, Knowledge Management, Data Mining, Networking, Parallel and Distributed Computing. He has to his credit 24 years of teaching and research experience. He is a senior member of International Association of Computer Science and Information Technology, Singapore and Active member of Computer Science Society of India, Computer Science Teachers Association, New York.