Probability-based imputation method for fuzzy cluster analysis of gene expression microarray data

Thanh Le, Tom Altman
Department of Computer Science and Engineering
University of Colorado Denver
Denver, CO 80217-3364, USA
lntmail@yahoo.com, tom.altman@ucdenver.edu

Katheleen J. Gardiner
Department of Pediatrics
Univ. of Colorado Denver Anschutz Medical Campus
Aurora, CO 80045, USA
katheleen.gardiner@ucdenver.edu

Abstract— Fuzzy clustering has been widely used for analysis of gene expression microarray data. However, most fuzzy clustering algorithms require complete datasets and, because of technical limitations, most microarray datasets have missing values. To address this problem, we present a new algorithm where genes are clustered using the Fuzzy C-Means algorithm, followed by approximating the fuzzy partition by a probabilistic data distribution model which is then used to estimate the missing values in the dataset. Using distribution-based approach, our method is most appropriate for datasets where the data are nonuniform. We show that our method outperforms six popular imputation algorithms on uniform and nonuniform artificial datasets as well as real datasets with unknown data distribution model.

Availability- The test datasets and the software are available online at http://ouray.ucdenver.edu/~tnle/fzpbi

Keywords- gene expression analysis; fuzzy c-means; missing data estimation; distribution-based imputation;

I. INTRODUCTION

Oligonucleotide and cDNA microarray technologies allow measurement of the expression levels of thousands of genes simultaneously. Microarray datasets are usually in form of matrices with rows containing the expression values of the genes and columns representing the different experimental conditions. Cluster analysis is used to discover relationships among genes by searching for patterns in the expression values. Gene expression datasets usually have missing values, due to experimental artifacts that may include weak or absent signals, spotting inconsistencies, scratches or dust on the slide or image corruption. In addition, if meta-analysis is done, the lack of an appropriate method to map probes onto genes across different microarray chip generations or microarray platforms can result in missing data.

Most clustering algorithms cannot be used with incomplete datasets because they use every attribute of each data object. Missing value correction is therefore essential as a pre-processing step and can be carried out by imputation. Timm et al. [13] and Kim et al. [7] showed that imputing missing values in each iteration of the clustering algorithm offers an advantage compared to missing value imputation during data pre-processing, because the former can apply the data distribution model estimated by the clustering algorithm to the missing value estimation. Garcia-Laencina et al. [4], by using empirical tests, showed that the K-Nearest Neighbors (KNN) algorithm imputation performed better than the Expectation Maximization (EM) algorithm imputation. Kim et al. [7] proposed a Fuzzy C Means (FCM) imputation method which performed better than both EM and KNN imputation methods. Di-Nuovo [2] also showed that FCM-based imputation methods provide better performance than EM imputation and Regression Imputation methods. Using FCM, Hathaway et al. [5] identified four different methods to solve the missing value problem. The whole-data strategy (WDS) and partial distance strategy (PDS) carry out the analysis using only available values, and some valuable information in the dataset may be therefore lost. The optimal complete strategy (OCS) estimates the missing values using membership-based centroids of the cluster prototypes and the nearest prototype strategy (NPS) simply uses the nearest prototype of each data object as an estimate of its missing values. Both the OCS and NPS methods impute the missing values during the FCM iterative process. Luo et al. [10] proposed an imputation method (FCMimp) that is similar to the OCS method but where the missing values are estimated only when the FCM algorithm is done. Mohammadi et al. [11] improved FCMimp by using a distance measure, which combines the internal distance measure with an external one using Gene Ontology (GO) terms. Kim et al. [7] proposed a method similar to OCS called Clustering Incomplete data using Alternating Optimization (CIAO). To avoid use of estimated values of the missing data that may be influenced by noise or improper imputation in the early iterations, CIAO includes a confidence degree, which starts at 0 and increases during the subsequent iterations of the imputation process.

A common limitation of existing methods for missing value imputation is that they use only the distance between the object and cluster centers and disregard information about cluster volume. In real-world datasets, data objects are not equally distributed among clusters and the clusters will differ in size. The distance from an object to the cluster center in a large cluster may therefore be larger than that of an object in a small cluster. If imputation of missing values and assignment of data objects are determined only on the basis of distance between the data object and cluster center, marginal objects of a large cluster may be incorrectly assigned to the immediately adjacent small cluster. This is illustrated in Fig.1 where, if a data object with missing values is in the gray rectangle, it may be incorrectly assigned to cluster 3 instead of cluster 2, and consequently, the

978-0-7695-4654-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ITNG.2012.159

IEEE computer society

missing attributes of this object will be estimated on the information of cluster 3 instead of cluster 2.

In this paper, we describe fzPBI, a Probability Based Imputation method for fuzzy cluster analysis of gene expression microarray data with missing values. Missing values are estimated using a probabilistic data distribution model derived from the possibility model generated by FCM.

II. METHODS

Cluster analysis decomposes a set of objects into clusters based on dissimilarity. In analysis of gene expression microarray datasets, we require the clustering to allow a single gene to belong to more than one cluster, because one gene may participate in multiple biological processes. FCM was chosen for this work because it provides both an effective mechanism for missing value imputation methods and allows genes to belong to multiple clusters.

A. Fuzzy C-Means algorithm (FCM)

Let $X = \{x_1, x_2,\ldots,x_n\} \in R^p$ be a set of data objects x_i, i=1..n, and for a given c, $2 \le c < n$, the Fuzzy C-Means algorithm (FCM) divides X into c clusters by minimizing the objective function:

$$J_m(U,V) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ki}^m d^2(x_i, v_k) \to \min, \qquad (1)$$

where $u_{ki} \in [0,1]\ \forall k,i$,

$$\sum_{k=1}^{c} u_{ki} = 1 \forall i\ ; \qquad (2)$$

m, $1 \le m$, is the fuzzifier factor; V, $V = \{v_1, v_2, \ldots, v_c\}$ is a set of c cluster centers; $U = \{u_{ki}\}_{i=1..n,k=1..c}$ is a partition matrix; and $d^2(.)$ denotes the Euclidean norm.

Minimizing J_m with respect to (2), we obtain an estimated model of U and V as:

$$u_{ki} = \left(\frac{1}{d^2(x_i, v_k)}\right)^{\frac{1}{m-1}} \Bigg/ \sum_{l=1}^{c}\left(\frac{1}{d^2(x_i, v_l)}\right)^{\frac{1}{m-1}}, \qquad (3)$$

$$v_k = \sum_{i=1}^{n} u_{ki}^m x_i \Big/ \sum_{i=1}^{n} u_{ki}^m . \qquad (4)$$

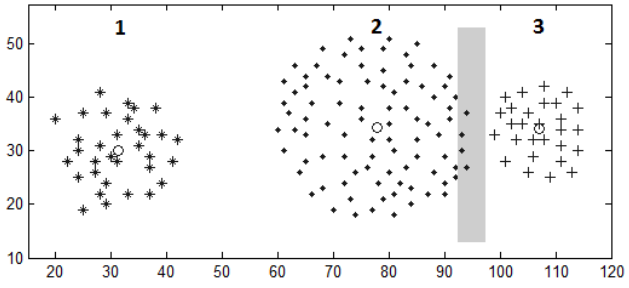

Figure 1. Artifical Dataset 2 (ASET2) with three clusters of different sizes

FCM can converge rapidly and provides soft partitions applicable to many real-world applications. However, FCM requires X to be complete.

B. Probability Based Imputation method for fuzzy cluster analysis (fzPBI)

The objective of fzPBI is to cluster a dataset X, that may be incomplete, into c clusters.

In a dataset with missing data, some attribute values of a data object x_i may be not observed. For example, $x_i = (x_{i1}, x_{i2}, ?, ?, x_{i5}, ?)$ has missing values corresponding to the third, fourth and sixth attributes. The missing values cause a problem with computing $d^2(.)$ between x_i and cluster centers. The FCM algorithm therefore cannot perform properly at x_i.

Let $X_W = \{x_i \in X \mid x_i \text{ is a complete data object}\}$, $X_P = \{x_{ij} \mid x_{ij} \ne ?\}$, and $X_M = \{x_{ij} \mid x_{ij}=?\}$. In fzPBI, X_M are estimated with respect to the optimization of the function J_m. The distance measure $d^2(.)$ is, therefore, defined as:

$$d^2(x_i, v_k) = \frac{p}{\sum_{j=1}^{p} w_j} \sum_{j=1}^{p} w_j \left(x_{ij} - v_{kj}\right)^2, \qquad (5)$$

where w_j indicates the contribution degree of the j^{th} attribute of x_i, x_{ij}, in the distance between x_i and v_k. If $x_{ij} \in X_M$, w_j increases with each iteration to avoid premature use of estimated values. Therefore, we define w_j as:

$$w_j = \begin{cases} 1 & x_{ij} \in X_P \\ t/T & x_{ij} \in X_M \end{cases}. \qquad (6)$$

where t is the iteration index, and T is the number of iterations; $0 \le t < T$.

To impute the values of X_M, we use the method of Le and Gardiner [8] to generate a probabilistic model of the data distribution using the fuzzy partition. This model is then used to impute the missing values. For each cluster v_k, k=1..c, a probability distribution $\{p_{ki}\}_{i=1..n}$ is derived from the possibility distribution $\{u_{ki}\}_{i=1..n}$. Then, the following statistics at v_k are computed:

$$\sigma_k = \sum_{i=1}^{n} p_{ki} \left\| x_i - v_k \right\|^2, \qquad (7)$$

$$P(v_k) = \sum_{i=1}^{n} P(x_i \mid v_k) \Bigg/ \sum_{l=1}^{c}\sum_{i=1}^{n} P(x_i \mid v_l), \qquad (8)$$

$$P(x_i \mid v_k) = \left((2\pi)^{1/n} \times \sigma_k \times e^{\frac{\|x_i - v_k\|^2}{2\sigma_k^2}} \right)^{-1}, \qquad (9)$$

where σ_k and $P(v_k)$ are the variance and the prior probability of v_k respectively; $P(x_i|v_k)$ indicates the conditional probability of x_i given v_k, i=1..n, k=1..c.

Let $P_k^m = \{p_{k1}^m, p_{k2}^m, ..., p_{kp}^m\}$ be a set of probabilities at v_k, where p_{kj}^m, j=1..p, indicates the probability that attribute j is missing in cluster k, defined as:

$$p_{kj}^m = \frac{\sum_{i=1}^{n} P(x_i, v_k)(1 - I_{ij})}{\sum_{i=1}^{n} P(x_i, v_k)}, \quad (10)$$

where

$$I_{ij} = \begin{cases} 1 & x_{ij} \in X_P \\ 0 & x_{ij} \in X_M \end{cases}.$$

Hence, the probability that a data object x_i has a missing attribute j in cluster k is

$$P(x_{M(ij)|k}) = P(v_k) p_{kj}^m \left[\prod_{t=1, t \neq j}^{p} p_{kt}^m \right] P(x_i \mid v_k). \quad (11)$$

Because each cluster v_k can be considered a component of the data distribution model of X, the estimated value of $x_{ij} \in X_M$ is computed as:

$$\hat{x}_{ij} = \sum_{k=1}^{c} P(x_{M(ij)|k}) v_k \Big/ \sum_{k=1}^{c} P(x_{M(ij)|k}). \quad (12)$$

To avoid an oscillation in the missing value estimation, estimated values are decreasingly used. In contrast to their usage in the clustering process, their contribution degree in distance measurement during the estimation process is defined as,

$$w'_j = \begin{cases} 1 & x_{ij} \in X_P \\ 1 - t/T & x_{ij} \in X_M \end{cases}. \quad (13)$$

fzPBI algorithm

Steps

1) t=0; initialize U_t randomly w.r.t (2).
2) Compute $\{d^2(x_i, v_k)\}_{i=1..n,k=1..c}$ using w_j(6).
3) Compute U^{t+1} using (3).
4) Compute V^{t+1} using (4).
5) If (t > T) or ($\left\| J_m^{t+1} - J_m^t \right\| < \varepsilon$) then Stop.
6) Compute $\{d^2(x_i, v_k)\}_{i=1..n,k=1..c}$ using w'_j(13).
7) Create a probabilistic data distribution model from the fuzzy partition using (7), (8) and (9).
8) Create a probabilistic model for X_M using (10) and (11).
9) Estimate X_M using (12). t=t+1. Go to Step 2.

The difference between fzPBI and the standard FCM algorithm is that fzPBI provides a method to discover the probabilistic model of the data distribution from the fuzzy partition and to apply the model to missing value imputation. For partition defuzzification, fzPBI uses the maximum membership degree of the data object. However, this method may be improper, particularly to marginal data objects of large clusters, as illustrated in Fig. 1. This specific problem requires more detailed analysis and will be the subject of future work.

III. EXPERIMENTAL RESULTS

Datasets

To evaluate the performance of fzPBI, we used two artificial datasets, ASET1 and ASET2, generated using an infinite mixture model method [15]. ASET1 has five well-separated clusters of similar size. ASET2 is more complex, containing three clusters that differ in size and density (Fig. 1). For the real datasets, we used the Iris and Wine datasets from the University of California Irvine (UCI) Machine Learning Repository [3]. For gene expression data, we used the RCNS dataset containing the expression levels of 112 genes over nine time points during the rat central nervous system development [14]; the Serum dataset containing 517 genes most responsive to treatment of human fibroblasts with serum [16]; the Yeast cell cycle dataset containing 384 genes [16]; and the Yeast-MIPS dataset [12], a subset of the Yeast cell cycle dataset containing 237 genes annotated with one of four specific functions (DNA synthesis and replication, organization of centrosome, nitrogen and sulphur metabolism, and ribosomal proteins). The classification structures in these datasets are presumed known.

Performance measures

We used two measures to evaluate algorithm performance. The first measure is the root mean square error (RMSE) between the true values and the imputed values, defined as:

$$RMSE = \sqrt{\frac{1}{n_i} \sum_{i=1}^{n_i} (x_i - \hat{x}_i)^2}, \quad (14)$$

where n_i is the number of missing values imputed. The second measure assesses the overall performance determined by the number of data objects with missing attributes that were misclassified. This assessment is done by comparing the cluster label of each data object with its actual class label. If the two match, there is no misclassification. If they do not match, then a misclassification has occurred.

We compared fzPBI with six popular imputation methods: PDS, OCS and NPS [5]; FCMimp [10]; CIAO [7]; and FCMGOimp [11]. For each dataset, we generated the missing data using different percentages of missing values. For the artificial datasets and the Iris, Wine and RCNS datasets, the value of the fuzzifier factor, m, was set to 2.0. For the Yeast, Yeast-MIPS and Serum datasets, m was set to 1.17 and 1.25 respectively, as in [1]. The number of clusters, c, was set to the known number of clusters. Each algorithm was run 5 times and the best result recorded. We repeated the

experiment 50 times and averaged the performance of each algorithm using both measures.

A. Artificial datasets

On ASET1, fzPBI had the lowest RMSEs across different scales of missing data (Fig. 2) and therefore performed better than the other methods. On ASET2, which contains clusters of different sizes, fzPBI again performed best (Fig. 3), although performance of the other methods was close.

B. Iris dataset

Applied to the Iris dataset, fzPBI and CIAO had the smallest RMSEs, although, as shown in Fig. 4, fzPBI performed marginally better. Table I shows that, compared with CIAO, fzPBI had a smaller number of misclassified objects across different scales of missing values.

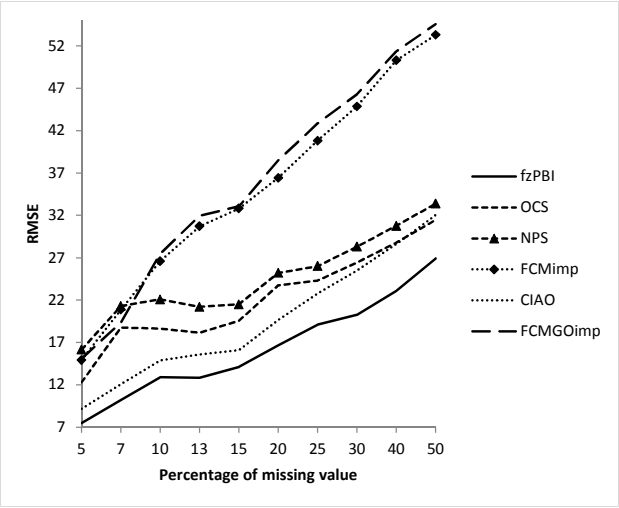

Figure 2. Average RMSE of 50 trials using an incomplete ASET1 dataset with different percentages of missing values

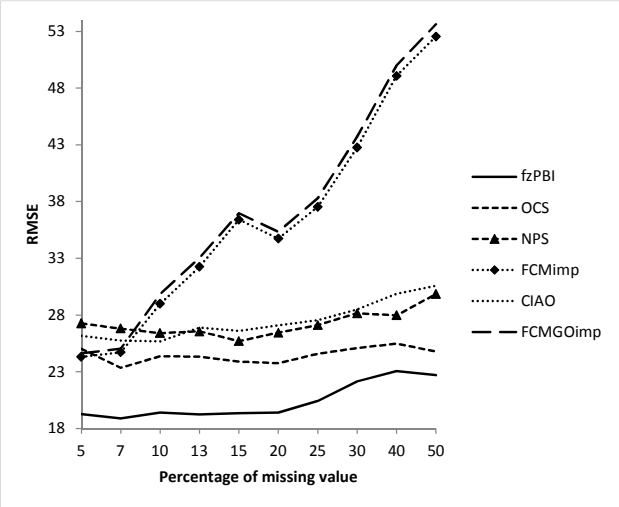

Figure 3. Average RMSE of 50 trials using an incomplete ASET2 dataset with different percentages of missing values

TABLE I. AVERAGE RESULTS OF 50 TRIALS USING AN INCOMPLETE IRIS DATASET WITH DIFFERENT PERCENTAGES (%) OF MISSING VALUE

%	Averaged #objects of misclassification						
	fzPBI	***PDS***	***OCS***	***NPS***	***FCMimp***	***CIAO***	***FCMGOimp***
5	15.9	18.7	16.0	16.2	18.4	15.9	18.0
7	15.9	12.6	16.3	16.8	13.0	15.9	13.2
10	15.9	10.0	16.3	17.7	9.8	16.0	9.5
13	16.0	9.1	17.1	19.0	9.5	16.2	9.0
15	16.1	12.8	18.4	20.7	12.2	16.3	9.9
20	15.9	20.6	19.8	22.9	20.3	16.0	23.7
25	16.2	31.9	20.5	24.8	30.8	16.1	32.8
30	16.7	37.9	26.2	30.9	37.9	16.7	38.9
40	16.7	49.3	30.8	37.9	50.8	17.0	57.8
50	21.2	56.5	42.6	52.3	57.1	21.8	64.1

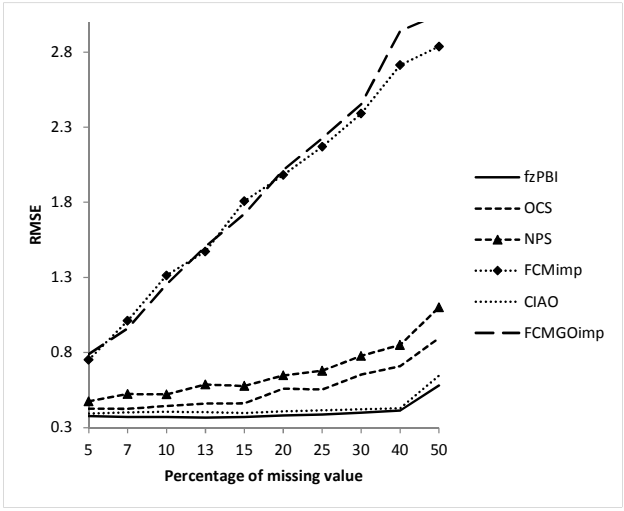

Figure 4. Average RMSE of 50 trials using an incomplete Iris dataset with different percentages of missing values

C. Wine dataset

On the Wine dataset, on average, both fzPBI and CIAO outperformed the other methods. fzPBI and CIAO performed equally on data with percentages of missing values of 10% and lower. However, fzPBI performed slightly better than CIAO on data with higher percentages of missing values. FCMimp and FCMGOimp performed better than other methods, when the percentage of missing values was small (Fig. 5). However, they performed worse than all other methods on datasets with high percentages of missing values.

D. RCNS dataset

On the RCNS dataset (Fig. 6), fzPBI outperformed other algorithms when the missing value percentage was 40% and lower. However, it performed marginally worse with a missing value percentage of 50%. This is because the RCNS dataset contains only 112 data points and becomes very sparse with high percentages of missing values; the data probability model was not properly determined, and fzDBI therefore could not perform correctly.

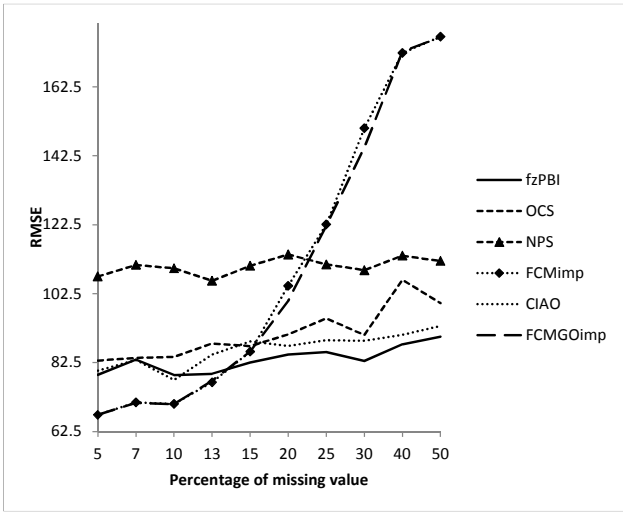

Figure 5. Average RMSE of 50 trials using an incomplete Wine dataset with different percentages of missing values

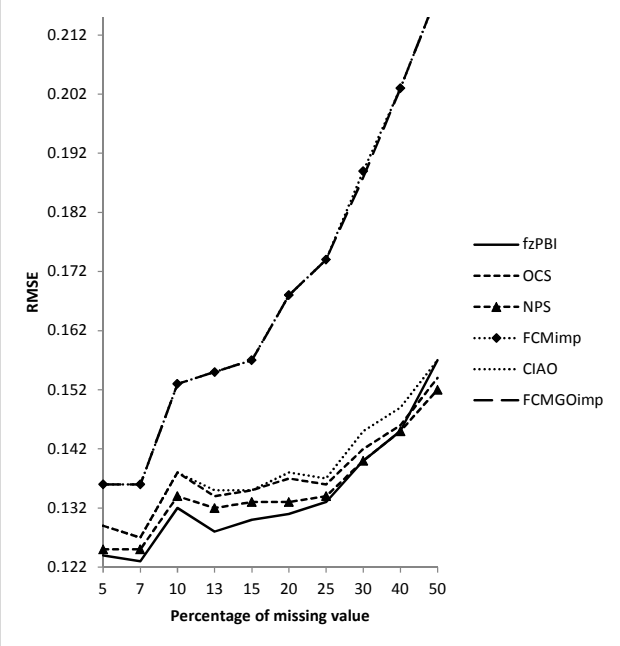

Figure 6. Average RMSE of 50 trials using an incomplete RCNS dataset with different percentages of missing values

E. Yeast and Yeast-MIPS datasets

fzPBI outperformed other methods on the Yeast and Yeast-MIPS datasets (Fig.s 7 and 8); FCMGOimp was run using the GO term-based distance measure [11]. However, FCMGOimp outperformed only FCMimp. This result is similar to that reported in [11]. Using GO terms to measure the distance between genes is interesting, however, a crisp distance measure of {0,1}, where 0 is the distance between a pair of genes having at least one GO term in common, does not help much with the problem of missing value imputation.

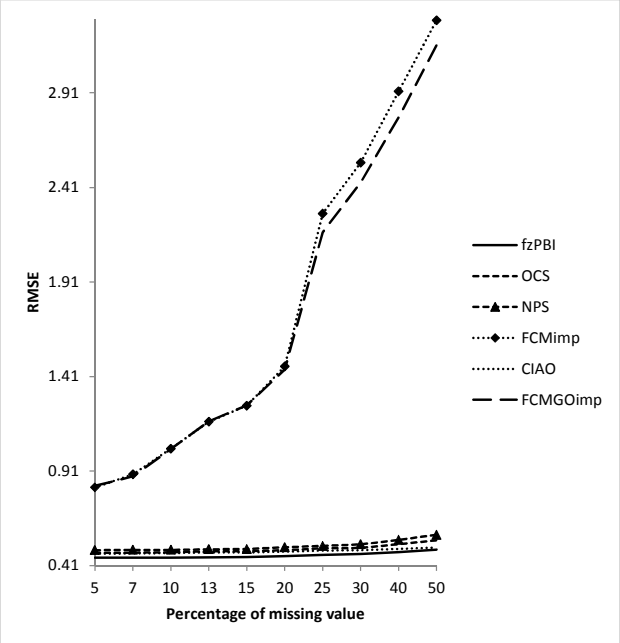

Figure 7. Average RMSE of 50 trials using an incomplete Yeast dataset with different percentages of missing values

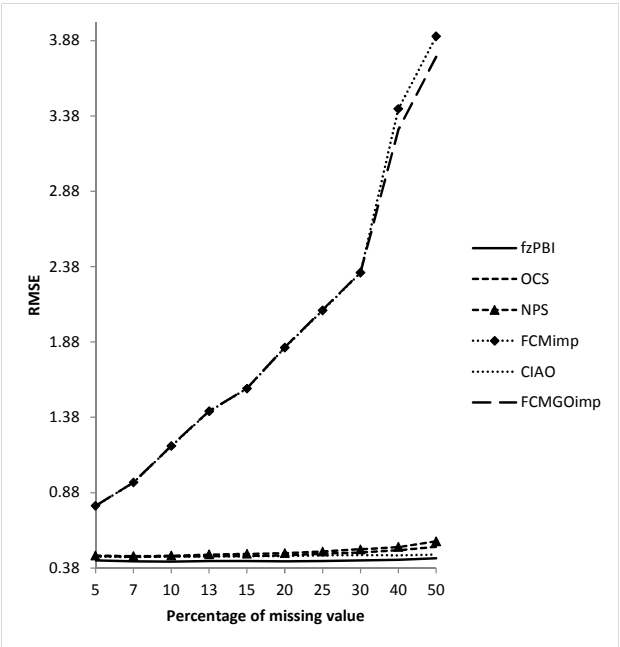

Figure 8. Average RMSE of 50 trials using an incomplete Yeast-MIPS dataset with different percentages of missing values

F. Serum dataset

fzPBI again outperformed other algorithms on Serum dataset with the percentage of missing value lower than 50% (Fig. 9). For the rest of the cases, fzPBI only performed worse than CIAO. On average, fzPBI performed best on the Serum dataset.

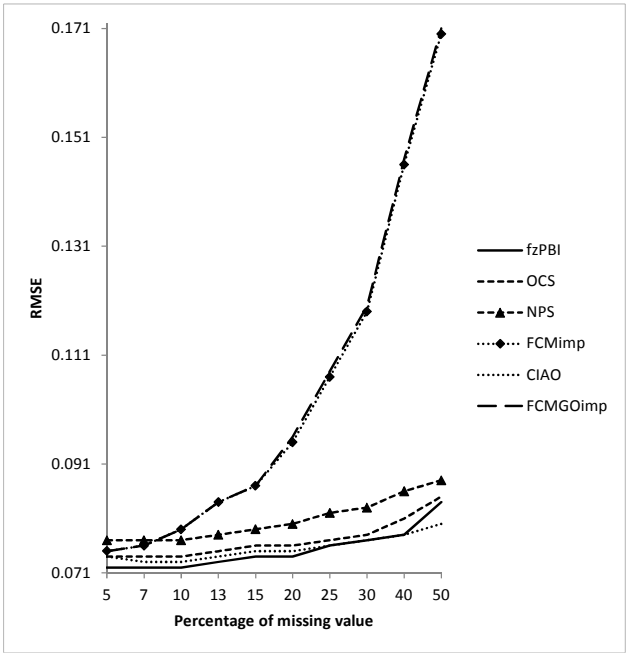

Figure 9. Average RMSE of 50 trials using an incomplete Serum dataset with different percentages of missing values

IV. CONCLUSIONS

We have presented fzPBI, a novel imputation method for fuzzy cluster analysis of gene expression microarray data with missing values. fzPBI outperformed other methods on both artificial and real datasets, particularly on datasets with clusters that differed in size. fzPBI is therefore appropriate for real-world datasets where the data densities are not uniformly distributed. In a similar study using this same data distribution based approach, we have previously demonstrated a density-based imputation method [9] that performed well on the same datasets as used here. This shows that the data distribution based approach is most appropriate for the problem of missing value imputation. In future work, we will exploit the relationships among GO terms at different levels to develop an external distance measure that effectively describes the biological distance between genes, and we will integrate this measure into fzPBI for a more powerful tool.

ACKNOWLEDGMENTS

This research was supported by the Vietnamese Ministry of Education and Training (to T.L.), and the Linda Crnic Institute for Down Syndrome and the National Institutes of Health HD056235 (to K.G.).

REFERENCES

[1] D. Dembele and P. Kastner, "Fuzzy C-means method for clustering microarray data," Bioinformatics, vol. 19, Jan. 2003, pp. 973-980, doi: 10.1093/bioinformatics/btg119.

[2] A.G Di-Nuovo, "Missing data analysis with fuzzy C-Means: A study of its application in a psychological scenario," Expert Syst. Appl., vol. 38, June 2011, pp. 6793–6797, doi:10.1016/j.eswa.2010.12.067.

[3] A. Frank and A. Asuncion, "Machine Learning Repository," [Online], http://archive.ics.uci.edu/ml, 2010.

[4] P.J. Garcia-Laencina, J.L. Sancho-Gomez, A.R. Figueiras-Vidal, "Pattern classification with missing data: a review," Neural Computing & Applications, vol. 19, March 2010, pp. 263–282, doi:10.1007/s00521-009-0295-6.

[5] R.J. Hathaway and J.C. Bezdek, "Fuzzy c-Means Clustering of Incomplete Data," Systems, Man and Cybernetics, vol. 31, Oct. 2001, pp. 735-744, doi:10.1109/3477.956035.

[6] V. R. Iyer et al., "The transcriptional program in the response of human fibroblasts to serum," Sci., Jan. 1999, pp. 83-87,doi: 10.1126/science.283.5398.83.

[7] D.W Kim, K.Y. Lee, K.H. Lee and D. Lee, "Towards clustering of incomplete microarray data without the use of imputation," Bioinformatics, vol. 23, Jan. 2007, pp. 107-113, doi: 10.1093/bioinformatics/btl555.

[8] T. Le and K.J. Gardiner, "A validation method for fuzzy clustering of gene expression data," Proc. Intl' Conf. Bioinformatics & Computational Biology (BIOCOMP'11), CSREA Press, July 2011, pp. 23-29.

[9] T. Le, T. Altman and K.J. Gardiner, "Density-based imputation method for fuzzy cluster analysis of gene expression microarray data," To appear in Proc. 4th Intl' Conf. Bioinformatics & Computational Biology (BICoB 2012), Las Vegas, Nevada, USA, Mar. 2012.

[10] J.W. Luo, T. Yang and Y. Wang, "Missing Value Estimation For Microarray Data Based On Fuzzy C-means Clustering," Proc. Intl' Conf. High-Performance Computing (HPCASIA'05), IEEE Press, Feb. 2006, pp. 11-16, doi:10.1109/HPCASIA.2005.53.

[11] A. Mohammadi and M.H Saraee, "Estimating Missing Value in Microarray Data Using Fuzzy Clustering and Gene Ontology," Proc. Intl' Conf. Bioinformatics and Biomedicine (BIBM'08), IEEE Press, Nov. 2008 pp. 382-385, doi: 10.1109/BIBM.2008.71.

[12] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church, "Systematic determination of genetic network architecture," Nat. Genet., vol. 22, 1999, pp. 281–285, doi: 10.1038/10343.

[13] H. Timm, C. Doring, R. Kruse, "Different approaches to fuzzy clustering of incomplete datasets," Intl' Journal of Approximate Reasoning, vol. 35, March 2004, pp.239-249, doi: 10.1016/j.ijar.2003.08.004.

[14] X. Wen et al., "Large-scale temporal gene expression mapping of central nervous system development," Proc. Intl' Conf. Natl Acad of Science, vol. 95, Natl Acad Press, Jan. 1998, pp. 334-339.

[15] L Xu and M.I. Jordan, "On convergence properties of the EM algorithm for Gaussian Mixtures," Neural Computation, vol. 8, Jan. 1996, pp. 129-151, doi:10.1162/neco.1996.8.1.129.

[16] K.Y. Yeung, D.R. Haynor, W. Ruzzo, "Validating clustering for gene expression data," Bioinformatics, vol. 17, April 2001, pp. 309–318, doi:10.1093/bioinformatics/17.4.309.