DENSITY-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 (FCM). The fuzzy partition obtained is then used to create a density-based fuzzy partition which is used with the FCM fuzzy partition to estimate the missing values in the dataset. We show that our method outperforms five popular imputation algorithms on both artificial and real datasets.*

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

Keywords *- microarray data; missing value estimation; fuzzy c-means; cluster density;*

1 INTRODUCTION

Oligonucleotide and cDNA microarray technology allows 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. [15] 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 it 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 Expectation Maximization (EM) algorithm imputation performs worse than the K-Nearest Neighbors (KNN) algorithm imputation. Kim et al. [7] proposed a Fuzzy C Means (FCM) imputation method which performed better than both EM and KNN imputation methods and Di-Nuovo [2] showed that FCM-based imputation methods also provide better performance than EM imputation and Regression Imputation. 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. [12] 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. [13] 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 NCS 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

This work was supported by the Linda Crnic Institute for Down Syndrome and the National Institutes of Health HD056235 (KG) and the Vietnamese Ministry of Education and Training (TL).

Thanh Le is a doctoral student in the Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO 80217-3364, USA (email: lntmail@yahoo.com).

information about cluster density and size. In real-world datasets, data objects are not equally distributed among clusters and the clusters will differ in density and 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 Figure 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, its missing attributes estimated on the information of cluster 3 instead of cluster 2.

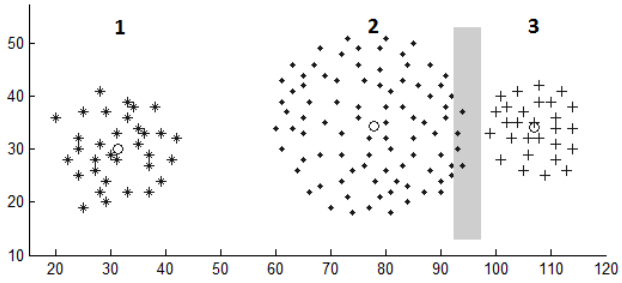

Figure 1: Artificial Dataset 2 (ASET2) with three clusters of different sizes

In this paper, we describe fzDBI, a Density Based Imputation method for fuzzy cluster analysis of gene expression microarray data with missing values. Missing values are estimated using both the fuzzy partition generated by FCM and the density-based fuzzy partition which, created based on the FCM fuzzy partition, describes cluster densities and volumes.

2 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.

2.1 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:

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

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

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

2.2 Density Based Imputation method for fuzzy cluster analysis (fzDBI)

The objective of fzDBI 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. FCM algorithm therefore cannot perform properly at x_i.

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

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

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; $0 \le t < T$.

Let $\theta = \{U,V\}$ be the fuzzy partition generated by FCM. θ is distance-based fuzzy partition which can describe the data density at the cluster centers. The accumulated density at the center of cluster v_k, k=1..c, is calculated as in (7). However, θ cannot describe properly the data density at every data point of X [11].

$$\mathrm{Acc}(v_k) = \sum_{i=1}^{n} u_{ki} . \tag{7}$$

To address this problem, we create a density-based fuzzy partition of X using the method of Le and Altman [8]. A *strong uniform fuzzy partition* is defined as

$$u'_{ki} = \left(e^{\frac{d^2(x_i,v_k)}{\sigma_k^2}} \right)^{-1} \Bigg/ \sum_{l=1}^{c} \left(e^{\frac{d^2(x_i,v_l)}{\sigma_l^2}} \right)^{-1}, \tag{8}$$

where σ_k, the variance of cluster v_k, k=1..c, is defined as

$$\sigma_k^2 = \sum_{i=1}^{n} P(x_i \mid v_k)\, d^2(x_i, v_k) \Big/ \sum_{i=1}^{n} P(x_i \mid v_k),$$

and $P(x_i|v_k)$, the conditional probability of x_k given v_i, i=1..n, k=1..c, is calculated based on the fuzzy partition using the method of Le and Gardiner [9]; the fzSC method [8] is then applied to compute the data density at every data point,

$$\mathrm{dens}(x_i) = \sum_{k=1}^{c} \mathrm{Acc}(v_k) \times u'_{ki}, \tag{9}$$

and to select the most dense data points, c_d, as cluster candidates, where c_d, $2 \le c_d \le \sqrt{n}$, is the desired number of clusters. The missing values of the k^{th} cluster candidate, k=1..c_d, if any exist, are estimated based on the FCM fuzzy partition,

$$\hat{v}_{kj} = \sum_{l=1}^{c} P(v_l \mid \hat{v}_k) \times v_{lj} \Big/ \sum_{l=1}^{c} P(v_l \mid \hat{v}_k). \tag{10}$$

Loquin et al. [11] suggested some *strong uniform fuzzy partition* models applicable to density estimation. In this study, using the Central Limit Theorem in cluster analysis [9], we proposed a model, as in (8), that gives the cluster most contributing to the data density at a data point the highest membership degree. This cluster will therefore play the most important role in imputation of missing values, if exist any, of the data point.

Let $\hat{V} = \{\hat{v}_1, \hat{v}_2, ... \hat{v}_{c_d}\}$ be a set of cluster candidates. We construct a density-based fuzzy partition $\hat{\theta}$, $\hat{\theta} = \{\hat{U}, \hat{V}\}$, where $\hat{U}$ is defined as

$$\hat{u}_{ki} = \sum_{l=1}^{c} P(v_l \mid \hat{v}_k) \times u'_{li} \Big/ \sum_{l=1}^{c} P(v_l \mid \hat{v}_k). \tag{11}$$

The estimated value of $x_{ij} \in X_M$ is computed based on both the FCM fuzzy partition, θ, and the density-based fuzzy partition, $\hat{\theta}$,

$$\hat{x}_{ij} = \alpha \sum_{k=1}^{c} u_{ki}^m v_{kj} \Big/ \sum_{k=1}^{c} u_{ki}^m + (1-\alpha) \sum_{k=1}^{c_d} \hat{u}_{ki}^m \hat{v}_{kj} \Big/ \sum_{k=1}^{c_d} \hat{u}_{ki}^m , \tag{12}$$

where α, $0<\alpha<1$, indicates the contribution level of each of the two fuzzy partition models, θ and $\hat{\theta}$, in the missing value imputation. We set $\alpha=0.5$ so that both the models contribute equally. To avoid an oscillation in the missing value estimation, estimated values are decreasingly used. In contrast with their usage in the clustering process, their contribution degree in distance measurement during the estimation process is defined as,

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

fzDBI algorithm

Steps

1) t = 0; c = 2×c_d.
2) Initialize U^t randomly w.r.t (2).
3) Compute $\{d^2(x_i, v_k)\}_{i=1..n,k=1..c}$ using w_j (6).
4) Compute U^{t+1} using (3).
5) Compute V^{t+1} using (4).
6) If ($t \ge T$) or ($\left\| J_m^{t+1} - J_m^t \right\| < \varepsilon$) then Stop.
7) Compute $\{d^2(x_i, v_k)\}_{i=1..n,k=1..c}$ using w'_j (13).
8) Create a strong uniform fuzzy partition using (8).
9) Create a density-based fuzzy partition of X, $\hat{\theta}$, using (9), (10) and (11).
10) Estimate X_M using (12). t=t+1. Go to Step 3.

The difference between fzDBI and the standard FCM algorithm is that, in addition to a distance-based fuzzy partition, fzDBI generates a density-based fuzzy partition which is used with the distance-based one in missing value imputation. For partition defuzzification, fzDBI 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 Figure 1. This problem requires a rigorous analysis which is the subject of future work.

3 EXPERIMENTAL RESULTS

Datasets

To evaluate the performance of fzDBI, we used two artificial datasets, ASET1 and ASET2, generated using an infinite mixture model method [17]. ASET1 has five well-separated clusters with similar size. ASET2 is more complex, containing three clusters that differ in size and density (Figure 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 [16]; the Serum dataset containing 517 genes most regulated underlying the response of human fibroblasts to serum [6]; the Yeast cell cycle dataset containing 384 genes [18]; the Yeast-MIPS dataset [14], a subset of the Yeast cell cycle dataset containing 237 genes annotated with one of four functions (DNA synthesis and replication, organization of centrosome, nitrogen and sulphur metabolism, and ribosomal proteins). The classification structures in these datasets are known.

Performance measure

We used the root mean square error (RMSE) between the true values and the imputed values, defined as in (14), to evaluate algorithm performance.

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

where n_i is the number of missing values imputed.

We compared fzDBI with five popular imputation methods: OCS and NPS [5]; FCMimp [12]; CIAO [7]; and FCMGOimp [13]. 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 desired number of clusters, c_d, 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 the performance measure.

Artificial datasets

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

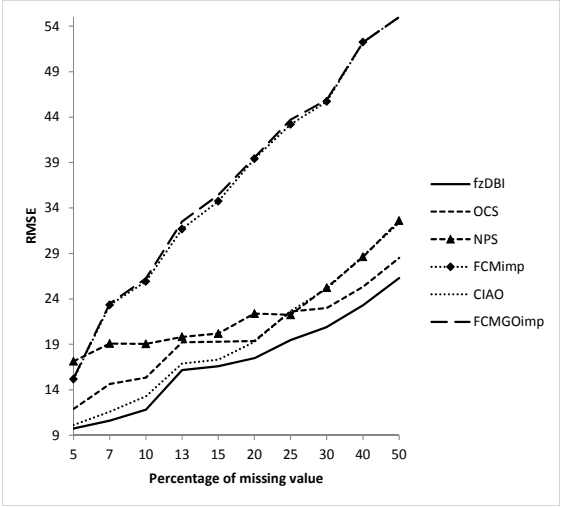

Figure 2: Average RMSE of 50 trials using an incomplete ASET1 dataset with different missing value percentages

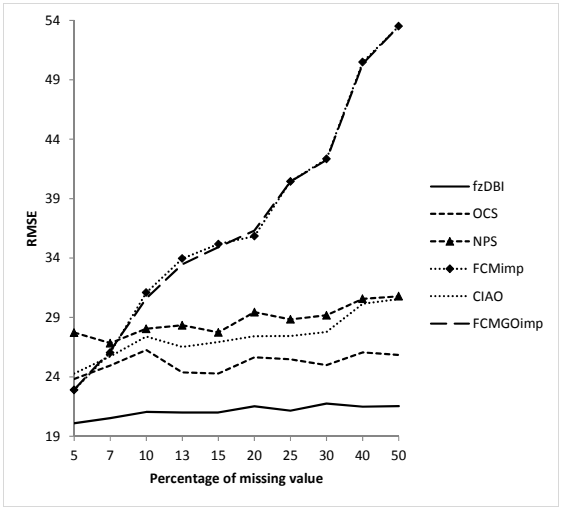

Figure 3: Average RMSE of 50 trials using an incomplete ASET2 dataset with different missing value percentages

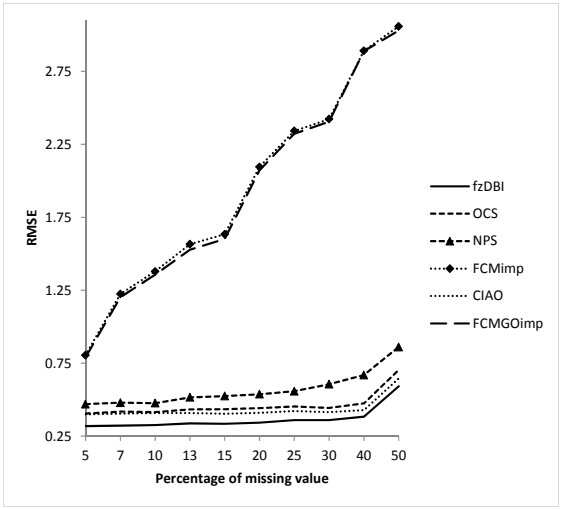

Figure 4: Average RMSE of 50 trials using an incomplete Iris dataset with different percentages of missing values

Iris dataset

Applied to the Iris dataset, fzDBI and CIAO had the smallest RMSEs, although, as shown in Figure 4, fzDBI performed marginally better.

Wine dataset

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

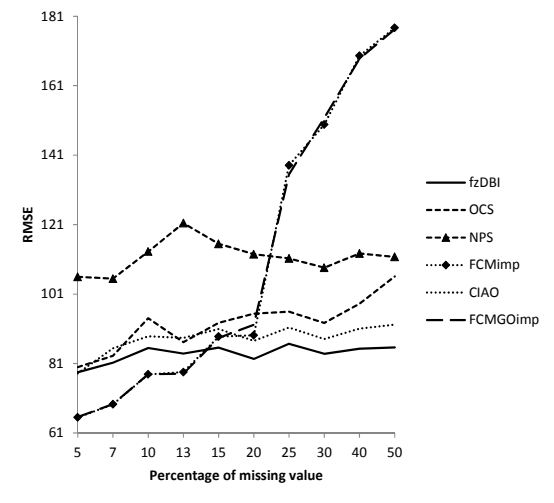

Figure 5: Average RMSE of 50 trials using an incomplete Wine dataset with different percentages of missing values

RCNS dataset

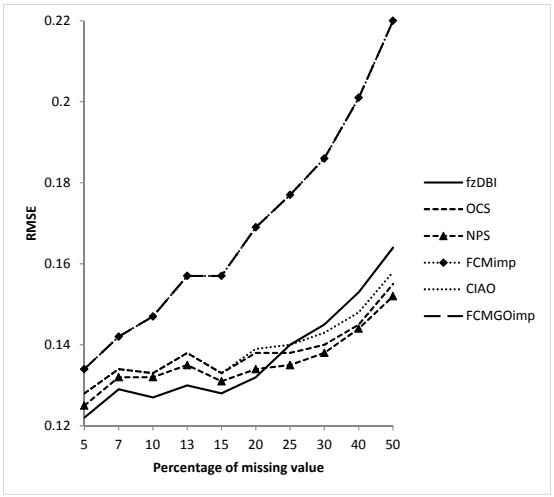

Figure 6: Average RMSE of 50 trials using an incomplete RCNS dataset with different percentages of missing values

On the RCNS dataset (Figure 6), fzDBI outperformed other algorithms when the missing value percentage was 20% and lower. It however performed worse with high missing value percentages. That is because the RCNS dataset, containing 112 data points, becomes very sparse with high percentages of missing values; fzDBI, working based on the data density, could not perform properly.

Yeast, Yeast-MIPS and Serum datasets

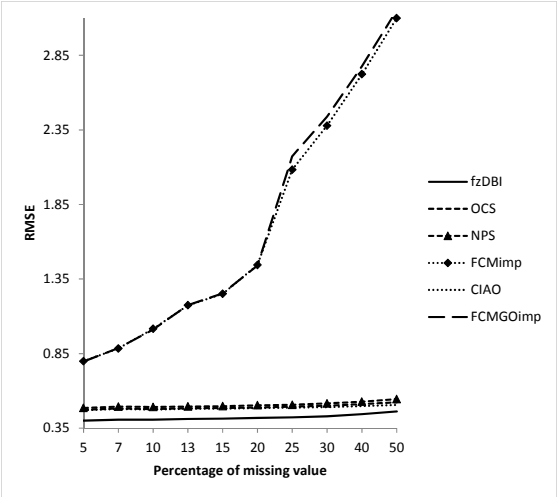

Figure 7: Average RMSE of 50 trials using an incomplete Yeast dataset with different percentages of missing values

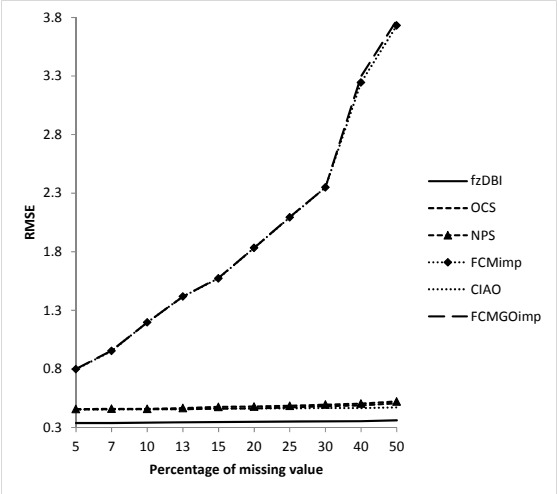

Figure 8: Average RMSE of 50 trials using an incomplete Yeast-MIPS dataset with different percentages of missing values

fzDBI outperformed other methods on the Yeast, Yeast-MIPS and Serum datasets (Figures 7, 8 and 9). For the Yeast and Yeast-MIPS datasets, FCMGOimp was run using the GO term-based distance measure [13]. However, FCMGOimp outperformed only FCMimp. This result is similar to that reported in [13]. 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.

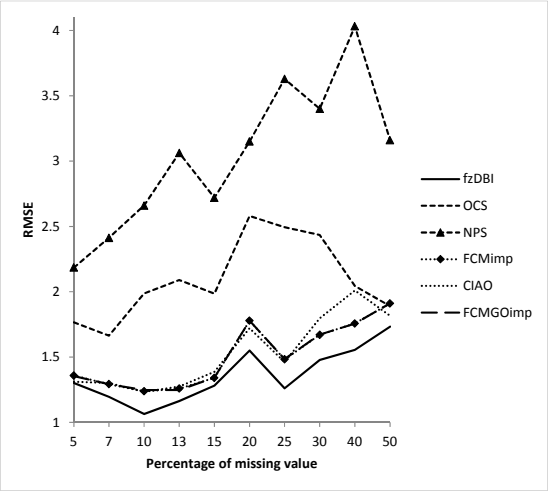

Figure 9: Average RMSE of 50 trials using an incomplete Serum dataset with different percentages of missing values

4 CONCLUSIONS

We have presented fzDBI, a novel density-based imputation method for fuzzy cluster analysis of gene expression microarray data with missing values. fzDBI outperformed other methods on both artificial and real datasets, particularly on datasets with clusters that differed in data density and size. fzDBI is therefore appropriate for real-world datasets in which the data densities are not uniformly distributed. Regarding this same problem, in a similar study using data distribution-based approach, we have also completed a probability-based imputation method [10] and achieved significant testing results on the same datasets with this study. 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 fzDBI for a more powerful tool.

5 REFERENCES

[1] D. Dembele and P. Kastner, "Fuzzy C-means method for clustering microarray data." Bioinformatics, Vol. 19, pp. 973-980, 2003.

[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, pp. 6793-6797, 2011.

[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 Comput Appl, Vol. 19, pp. 263-282, 2010.

[5] R.J. Hathaway and J.C. Bezdek, "Fuzzy c-Means Clustering of Incomplete Data," Systems, Man and Cybernetics, Vol. 31, pp. 735-744, 2001.

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

[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, pp. 107-113, 2007.

[8] T. Le and T. Altman, "A new initialization method for the Fuzzy C-Means Algorithm using Fuzzy Subtractive Clustering," Proc. Intl' Conf. on Information and Knowledge Engineering, pp. 144-150, Las Vegas, Nevada, USA, 2011.

[9] T. Le and J.K. Gardiner, "A validation method for fuzzy clustering of gene expression data," Proc. Intl' Conf. on Bioinformatics & Computational Biology, Vol. I, pp. 23-29, Las Vegas, Nevada, USA, 2011.

[10] T. Le, T. Altman and J.K. Gardiner, "Probability-based imputation method for fuzzy cluster analysis of gene expression microarray data," To appear in Proc. Intl' Conf. on Information Technology-New Generations, Las Vegas, Nevada, USA, 2012.

[11] K. Loquin and O. Strauss, "Histogram density estimators based upon a fuzzy partition," Statistics and Probability Letters, Vol. 78, pp. 1863-1868, 2008.

[12] J.W. Luo, T. Yang, Y. Wang, "Missing Value Estimation For Microarray Data Based On Fuzzy C-means Clustering," Proc. Intl' Conf. on High-Perform Comput, Changsha China, 2005.

[13] A. Mohammadi and M.H Saraee, "Estimating Missing Value in Microarray Data Using Fuzzy Clustering and Gene Ontology," Proc. Intl' Conf. on Bioinformatics and Biomedicine, pp. 382-385, Washington, DC, USA. 2008.

[14] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church, "Systematic determination of genetic network architecture," Nat. Genet. 22, 281-285, 1999.

[15] H. Timm, C. Doring, R. Kruse, "Different approaches to fuzzy clustering of incomplete datasets," Intl' Journal of Approximate Reasoning, Vol. 35, pp.239-249, 2004.

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

[17] L Xu and M.I. Jordan, "On convergence properties of the EM algorithm for Gaussian Mixtures." Neural Computation, Vol. 8, pp. 409-1431, 1996.

[18] K.Y. Yeung, D.R. Haynor, W. Ruzzo, "Validating clustering for gene expression data," Bioinformatics, Vol. 17, pp. 309-318, 2001.