

Estimating Missing Value in Microarray Gene Expression Data Using Fuzzy Similarity Measure

Amit Paul

Computer Science and Engineering Department
St. Thomas College of Engineering and Technology
khidirpore, India
Email: amitpaul83@gmail.com

Jaya Sil

Computer Science and Technology Department
Bengal Engineering and Science University
Shibpur, India
Email: js@cs.becs.ac.in

Abstract—Microarray experiments usually generate data sets with multiple missing value due to several reasons. In the paper a robust method has been proposed to estimate the missing value of microarray experimental data. Missing values are imputed using fuzzy similarity measure by identifying the genes having similar characteristics to that of the gene with missing values. In this approach, biological knowledge of the gene is extracted using fuzzy relation and based on that knowledge, missing value is predicted and optimized. The estimation accuracy of the proposed method is compared with the existing K-nearest neighbour (KNN) based missing value imputing method. The result demonstrates that the proposed method outperforms the KNN based method.

Index Terms—Microarray gene expression, missing value, fuzzy similarity, biological knowledge

I. INTRODUCTION

Microarray technology, a high throughput technology [1] evaluates the expression of thousands of genes simultaneously [2]. A microarray experiment typically assesses a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags) under multiple conditions. These conditions may be a time series during a biological process or a collection of different tissue samples (e.g., normal versus cancerous tissues). Data obtained from a microarray experiment is represented by a real-valued gene expression matrix, where the rows form the expression patterns of genes, columns represent expression profiles of the samples and each cell is the measured expression level of gene samples.

Usually, data from microarray experiments contain missing value due to different reasons including dust or scratches on the slide, error in experiments, image corruption and insufficient resolution. Typically 1-50% of data are missing [3], affecting up to 95% of the genes. Missing value is a predominant problem for the analysis of microarray data, which needs attention in order to obtain accurate results from microarray data analysis [2], [4]. Therefore, as a preprocessing step missing value estimation is essential. Existing algorithms require a full dataset for statistical analysis of microarray data [4] because the underlying statistical methodology is based on balanced datasets. There are many approaches to deal with missing

value, including:

- a) Repeating the experiment [5]
- b) Ignoring objects containing missing value [6]
- c) Estimating the missing value

The first approach is expensive and time consuming while the second approach usually lost too much useful information and may bias the results if the remaining cases don't represent the entire sample properly. The third approach can be subdivided into two groups. The missing value estimation methods of the first group substitute the missing values by a global constant such as 'zero' [6] or by the average of the available gene expression values [7] resulting distortion in relationships among the variables. Correlation structure is considered in the second group in two steps. At first, genes similar to the gene with missing value(s) are identified and then the missing values are predicted with the help of the observed values of the identified genes. The widely used weighted K -nearest neighbor imputation (KNN impute) procedure [8] reconstructs the missing values using a weighted average of K most similar genes [9]. These methods have better performance though the estimation ability largely depends on the selection of K parameter. There is no theoretical way to determine the value of K appropriately but to specify by the user. In cluster-based algorithms [10], [11] determination of missing values have been proposed where number of cluster either predetermined or the user have to specify that. Moreover, since the accuracy of the clustering algorithms depends on number of clusters, its selection becomes very crucial. The limitation of the existing methods is due to the fact that the estimation is based solely on the gene expression data ignoring biological information, which is inherent in the gene dataset.

Fig. 1 shows three genes g_1 , g_2 , and g_3 on seventeen experimental conditions c_j for $j = 1, \dots, 17$ where the y -axis shows the expression level of a gene g_i on a particular condition c_j , plotted along x axis. Based on a distance function g_1 and g_2 are clustered together. However, from the biological viewpoint, it is more appropriate to cluster g_1 and g_3 because g_1 and g_3 exhibit similar tendency along a subset of conditions. It is therefore, concluded [12]-[14] that strong correlations may

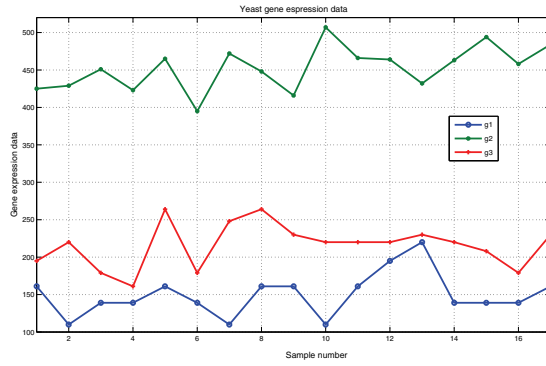


Fig. 1. Three genes with seventeen experimental conditions.

exist among a set of genes, even if their spatial localities are far apart from each other. Pattern-based and tendency-based biclustering algorithms were proposed [15] to overcome the conventional constraints of spatial proximity and able to discover clusters in all subspaces by pattern/tendency similarities. However, a pattern-based model is only capable of identifying positive correlated genes if and only if some linear transformation relationships such as shifting [14], [16], [17] and scaling [18] exist among their expression levels. The mechanism of gene regulation is complicated and becomes oversimplified if one assumes that only linear relationships exist in coregulated gene profiles [19]. Order Preserving Cluster (OP-CLUSTER) [12] and Tendency Preserving cluster (TP-CLUSTER) [13] are tendency based clustering algorithms adopt a model for efficient discovery of genes. Here, the gene expression levels synchronously rise and fall in a subspace while at the same time do not impose any particular magnitude transformation relationships among the genes. However, recent research work [19], [20] provides a greater diversity of correlations among coregulated genes such as genes with a negative pattern or inverted behavior, where existing pattern-based and tendency-based algorithms lost their significance.

In the paper, a robust method has been proposed to estimate the missing value of microarray experimental data by considering both positive and negative correlated genes. Missing values are imputed by identifying the genes which have similar characteristics to that of the gene with missing value, determined by soft class membership value using fuzzy similarity measure. The paper briefly addresses the approaches for finding related genes followed by the discussions of necessities for finding co-relational gene, with support from real examples.

The rest of the paper is organized as follows. The proposed method is described in section II. Experimental results are presented in section III. Concluding remarks and suggestions for future work are given in section IV.

II. METHODOLOGY

Clustering is an unsupervised learning technique that aims at partitioning a given set of objects into subgroups or clusters

based on the similarity of objects. It has been observed that each gene may be involved in more than one biological activity, therefore, hard clustering which assign each gene to a single cluster fails to express the total characteristics of the genes [21]. In the paper, first gene expression dataset is normalized and then clustered using fuzzy c means clustering algorithm [22] that measures similarity between genes. The belongingness of a single gene into multiple clusters is determined with varied degree of membership value obtained using clustering analysis. The proposed method determines the functional similarity between the genes inherent in the sample values in order to acquire more biologically interpretable clusters.

A. Pattern Similarity Matching (PSM) Algorithm

Few terminologies are defined below for understanding the proposed PSM algorithm.

Normalized gene expression ($NGE_{(i,j)}$): Difference between the j^{th} sample value of each i^{th} gene_expression and mean of the sample values of the corresponding gene_mean represent zero mean normalize gene expression, defined in (1). The available sample values of genes are normalized and the gene expression matrix is reconstructed with the new gene expression data (*Normalized gene expression*). This represent zero mean relational activity level of that gene for the sample.

$$NGE_{(i,j)} = gene_expression_{(i,j)} - gene_mean_{(i,j)} \quad (1)$$

Activation level ($Al_{(i)}$): Distribution of gene expression data in two-dimension space reveals the fact that several genes are highly active compare to others. Initialize the activation level of each gene to zero. Increase the activation level of gene i by one when its present sample, say j having higher value than its previous sample, as defined in (2).

$$Al_{(i)} = Al_{(i)} + 1 \quad \text{if } NGE_{(i,j)} > NGE_{(i,j-1)} \quad (2)$$

Depending on the activity of gene i , its class membership value $Mfl_{(i)}$ is evaluated using the concept of fuzzy similarity measure, defined in (3) that represent its degree of belongingness in a particular class.

$$Mfl_{(i)} = \frac{Al_{(i)}}{gs} \quad (3)$$

where gs is the number of samples in gene i without any missing value.

Missing gene list(misg): The list contains genes having one or more missing sample value.

Not A Number (NaN): Not a Number (NaN)[23] represents fractional sample value of a gene. NaN's are represented by a bit pattern with an exponent of all 1s with non-zero fraction value.

Gene expression data represents the logarithmic value of gene's protein production under a particular time interval. Generally, each gene expression data is distinct, so normalization

is required to find out the similar active genes. The normalized value of a particular sample is basically information about the gene's protein production, typically the zero mean, depicted in (1). The PSM algorithm calculates similarity of the genes based on their class membership value (Mfl) as given in equation (3). If a set of genes ($i = 1 \dots n$) have similar activity level of protein production with a gene having missing value (l) then a particular sample position among the similar genes is selected using (4) to impute the missing value at k^{th} sample position of l^{th} gene.

$$Optv = \min_{(\forall i, \forall j)} (|gene_expression(i, j) - gene_expression(i, k)|) \quad (4)$$

$Optv$ is the difference between the sample value at missing position k with other sample positions of the respective gene belonging to the same cluster. Minimum difference occurs say, at sample position s of a particular gene. Now the sample value at position s of the missing gene (l) is used to impute the missing value at position k .

Missing gene expression value can be estimated from k^{th} sample position of missing gene l .

B. Procedure:

Input: Microarray gene expression data, with some genes' (object) sample (attribute) value missing.

Output: Microarray gene expression data, with some artificial data impute.

Step1: $mean_gene := \text{mean}(gene_array)$

Step2: $misg := \text{find_missing_gene}(gene)$

Step3: for all i gene in $gene_array$ do

for all j sample of the gene do

$NGE_{(i,j)} := gene_array_{(i,j)} - mean_gene_{(i)}$

if $(NGE_{(i,(j-1))} > NGE_{(i,j)})$ then

$Al_{(i)} := Al_{(i)} + 1$

end if

end j^{th} loop

$Mfl_{(i)} := \frac{Al_{(i)}}{gs}$

end i^{th} loop

Step4: for all l gene having missing value in $misg$ do
 $smi_Mfl_{(l)} := \text{fuzzy similar gene } (Mfl_{(l)})$

$miss := \text{find missing sample}(gene_array(misg(l)))$

for i^{th} gene in $smi_Mfl_{(l)}$ do

for all k missing sample in $miss$ do

for all j sample in $smi_Mfl_{(l,i)}$

without NaN gene expression for that sample in l^{th} and i^{th} both gene

$$Optv := |gene_array_{(smi_Mfl_{(l,i),j})} - gene_array_{(smi_Mfl_{(l,i),k})}|$$

if $Optv$ is minimum then

$selected_sample := j^{th}$

Step5:

$$gene_array_{(misg(l),k)} := gene_array_{(misg(l),selected_sample)}$$

The following example illustrates the above procedure and summarized in tables I, II, III, IV and V by considering twelve genes, each having six samples.

TABLE I
AFTER RANDOMLY DELETION OF DATA FROM ORIGINAL MICROARRAY GENE EXPRESSION MATRIX

Gene id	S_1	S_2	S_3	S_4	S_5	S_6
G_1	7.5	7.58	7.66	7.68	7.01	7.47
G_2	80.2	80.7	80.3	NaN	81	82
G_3	8	8.9	9	9.5	8	8.5
G_4	80	82	84	88	81	79
G_5	2.5	3	2.7	3.2	4	3.5
G_6	3	3	2.7	3.2	3.5	3.5
G_7	72.4	73	62	61	60	58
G_8	82.3	83	82.7	80.2	80	80.5
G_9	92.5	92	91.7	93.2	91	90.8
G_{10}	32	31.6	31.7	31.8	31	30
G_{11}	50	49	48	48.5	49	50
G_{12}	62	62.8	62.1	62.5	62.1	62

TABLE II
MEAN OF EACH GENE IN MICROARRAY GENE EXPRESSION MATRIX

Gene id	Mean
G_1	7.48
G_2	80.84
G_3	8.43
G_4	82.33
G_5	3.15
G_6	3.15
G_7	64.4
G_8	81.45
G_9	91.87
G_{10}	31.35
G_{11}	49.08
G_{12}	62.25

According to the PSM algorithm, $Optv$ value (tableV) of all genes (tableI) that are similar to gene with missing value i.e. G_2 are calculated. The minimum value ($|G_1, S_4 - G_1, S_3| = 0.02$) occurs at sample position 3. Now impute (G_2, S_4) using 80.3 from (G_2, S_3) to remove missing value in the original microarray gene expression matrix (table I) .

TABLE III
RECONSTRUCTION OF MICROARRAY GENE EXPRESSION MATRIX AFTER
NORMALIZATION OF DATA

Gene id	S_1	S_2	S_3	S_4	S_5	S_6
G_1	0.02	0.10	0.18	0.20	-0.47	-0.01
G_2	-0.64	-0.14	-0.54	NaN	0.16	1.16
G_3	-0.43	0.47	0.57	1.07	-0.43	0.07
G_4	-2.33	-0.33	1.67	5.67	-1.33	3.33
G_5	-0.65	-0.15	-0.45	0.05	0.85	0.35
G_6	-0.15	-0.15	-0.45	0.05	0.35	0.35
G_7	8.00	8.60	-2.40	-3.40	-4.40	-6.40
G_8	0.85	1.55	1.25	-1.25	-1.45	-0.95
G_9	0.63	0.13	-0.167	1.33	-0.867	-1.067
G_{10}	0.65	0.25	0.35	0.45	-0.35	-1.35
G_{11}	0.92	-0.08	-1.08	-0.58	-0.08	0.92
G_{12}	-0.25	0.55	-0.15	0.25	-0.15	-0.25

TABLE IV
ACTIVITY LEVEL AND Mfl OF EACH GENE IN MICROARRAY GENE
EXPRESSION MATRIX

Gene id	Activity level	Mfl
G_1	4	0.67
G_2	3	0.6
G_3	4	0.67
G_4	4	0.67
G_5	3	0.5
G_6	2	0.33
G_7	1	0.167
G_8	1	0.167
G_9	1	0.167
G_{10}	2	0.33
G_{11}	2	0.33
G_{12}	2	0.33

The following example illustrates the above procedure with multiple missing values and summarized in tables VI, VII, VIII, IX, X and XI by considering fourteen genes, each having six samples.

Using (4) $Optv$ value (tableX) of all genes (tableVI) that are similar to gene with missing value i.e. G_1 are calculated. Minimum value ($|G_3, S_4 - G_3, S_3| = 0.5$) occurs at sample position 3, as given in table X. Now impute (G_1, S_4), which had missing value in the original microarray gene expression matrix (table VI) using 7.77 from (G_1, S_3). Similarly (G_2, S_3) is imputed from (G_2, S_2) considering similar gene G_4 . It has been observed that G_3 with multiple missing value is similar to G_1 and minimum $Optv$ value is attained in (G_1, S_6) from table XI. But (G_3, S_6) is NaN, so search for next minimum $Optv$ value, occurred at (G_1, S_1) and (G_3, S_5) is imputed using (G_3, S_1).

TABLE V
 $Optv$ FOR GENE G_1

Sample id	$Optv$
S_1	0.18
S_2	0.1
S_3	0.02
S_5	0.67
S_6	0.21

TABLE VI
AFTER RANDOMLY DELETION OF DATA FROM ORIGINAL MICROARRAY
GENE EXPRESSION MATRIX

Gene id	S_1	S_2	S_3	S_4	S_5	S_6
G_1	7.38	7.45	7.77	NaN	7	7.2
G_2	40.65	40	NaN	41	41.20	41.80
G_3	34	35.7	35	35.2	NaN	NaN
G_4	34	34.8	34.2	34	33	34
G_5	8	8.9	9	9.5	8	8.5
G_6	80	82	84	88	81	79
G_7	2.5	3	2.7	3.2	4	3.5
G_8	3	3	2.7	3.2	3.5	3.5
G_9	72.4	73	62	61	60	58
G_{10}	82.3	83	82.7	80.2	80	80.5
G_{11}	92.5	92	91.7	93.2	91	90.8
G_{12}	32	31.6	31.7	31.8	31	30
G_{13}	50	49	48	48.5	49	50
G_{14}	62	62.8	62.1	62.5	62.1	62

III. RESULTS AND COMPARISONS

A. Data set

In this paper, the performance of each method is evaluated using [24]. RNA was obtained from histologically normal bronchial epithelium of smokers during time of clinical bronchoscopy from relatively accessible airway tissue. Gene expression data from smokers with lung cancer was compared with samples of smokers without lung cancer.

B. Measurements for performance

The gene expression matrix obtained from the microarray experiment is modified by replacing all the rows containing missing values with randomly generated artificial missing

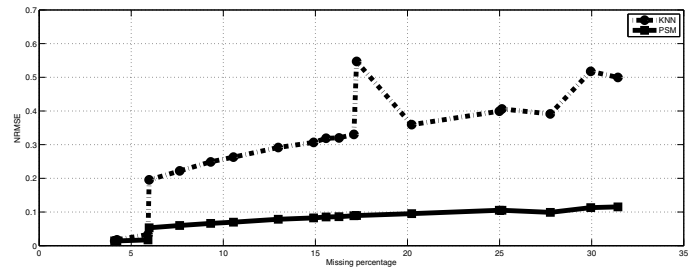


Fig. 2. Comparison of NRMSE value with PSM and KNN for GSE4115 data set

TABLE VII
MEAN OF EACH GENE IN MICROARRAY GENE EXPRESSION MATRIX

Gene id	Mean
G_1	7.36
G_2	40.93
G_3	34.975
G_4	34
G_5	8.43
G_6	82.33
G_7	3.15
G_8	3.15
G_9	64.4
G_{10}	81.45
G_{11}	91.87
G_{12}	31.35
G_{13}	49.08
G_{14}	62.25

TABLE VIII
RECONSTRUCTION OF MICROARRAY GENE EXPRESSION MATRIX AFTER NORMALIZATION OF DATA

Gene id	S_1	S_2	S_3	S_4	S_5	S_6
G_1	0.02	0.09	0.41	NaN	-0.36	-0.16
G_2	-0.28	-0.93	NaN	0.07	0.27	0.87
G_3	-0.975	0.725	0.025	0.225	NaN	NaN
G_4	0	0.8	0.2	0	-1	0
G_5	-0.43	0.47	0.57	1.07	-0.43	0.07
G_6	-2.33	-0.33	1.67	5.67	-1.33	3.33
G_7	-0.65	-0.15	-0.45	0.05	0.85	0.35
G_8	-0.15	-0.15	-0.45	0.05	0.35	0.35
G_9	8.00	8.60	-2.40	-3.40	-4.40	-6.40
G_{10}	0.85	1.55	1.25	-1.25	-1.45	-0.95
G_{11}	0.63	0.13	-0.167	1.33	-0.867	-1.067
G_{12}	0.65	0.25	0.35	0.45	-0.35	-1.35
G_{13}	0.92	-0.08	-1.08	-0.58	-0.08	0.92
G_{14}	-0.25	0.55	-0.15	0.25	-0.15	-0.25

values. The performance is measured by Normalise Root Mean Square Error (NRMSE) evaluated for error detection of Y_{guess} , where Y_{guess} is estimated by the proposed algorithm and Y_{answer} is the original microarray gene expression value.

$$NRMSE = \frac{\sqrt{(\text{mean}[(Y_{guess} - Y_{answer})^2])}}{\text{Var}(Y_{answer})} \quad (5)$$

Where $\text{Var}(x)$ returns the variance of matrix x and $\text{mean}(y)$ mean of the y matrix.

In Figure 2 and 3 comparison of NRMSE value for GSE4115 and GDS2771 series data are shown by applying PSM and KNN algorithms, respectively. Estimated sample value based on PSM impute method is closer to experimental data. KNN impute method estimates zero value if the sample value at the same position of k nearest neighbor genes is missing. On the other hand, if every gene consists missing

TABLE IX
ACTIVITY LEVEL AND Mfl OF EACH GENE IN MICROARRAY GENE EXPRESSION MATRIX

Gene id	Activity level	Mfl
G_1	3	0.6
G_2	2	0.4
G_3	2	0.5
G_4	2	0.33
G_5	4	0.67
G_6	4	0.67
G_7	3	0.5
G_8	2	0.33
G_9	1	0.167
G_{10}	1	0.167
G_{11}	1	0.167
G_{12}	2	0.33
G_{13}	2	0.33
G_{14}	2	0.33

TABLE X
 $Optv$ OF G_3 FOR GENE G_1

Sample id	$Optv$
S_1	1.7
S_2	0.7
S_3	0.5
S_5	NaN
S_6	NaN

value then KNN fails to impute any such missing value. However, the proposed PSM method able to impute 100% missing value in several passes varying with different rate of success. Keeping the missing percentage same with respect to microarray data, increase of number of missing value resulting rapid increase of NRMSE value in case of KNN while in case of PSM the rate is significantly less.

The PSM impute method shows comparable results with the previously developed methods in present research. The outstanding estimation ability of this impute method is partly due to the use of the most missing value information by incorporating fuzzy relation scheme. The results show that the estimating sample values obtained by the proposed PSM

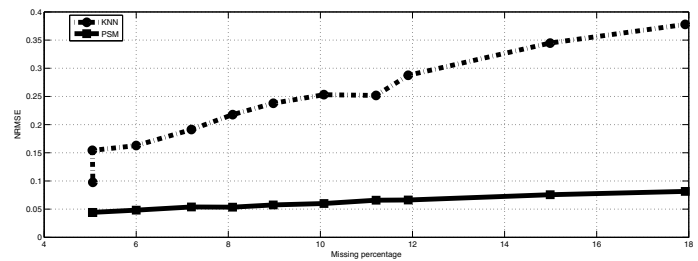


Fig. 3. Comparison of NRMSE value with PSM and KNN for GDS2771 data set

TABLE XI
Optv OF G_1 FOR GENE G_3

Sample id	Optv
S_1	0.38
S_2	0.45
S_3	0.77
S_5	NaN
S_6	0.2

impute method of gene expression data for different genes are less deviated with respect to the microarray data. KNN method depends on how far sample value of K number of genes are closed with the missing value ignoring the characteristic pattern of genes, which are might be different. But the PSM algorithm depends on the characteristic pattern, and here K varies depending on characteristics of genes. KNN impute method find out K nearest gene for each missing value from the remaining gene. Time complexity is high in KNN method due to large number of comparisons, whereas in PSM based impute method time complexity depends on number of unique active level. Therefore, application of PSM algorithm provides better result in less time than KNN impute method in microarray data set consisting of genes with sample values not very close to each other.

IV. CONCLUSION

The proposed PSM method demonstrates better results in comparisons with the existing method. Here biological knowledge has been exploited to determine similar genes for prediction of missing values unlike the previous works where distance measures were used for missing value prediction, ignoring the inherent knowledge of the genes. The missing value in a gene is predicted from the available value of the same gene, resulting least error in estimation of sample value. The optimized missing value thus provides near exact knowledge for important application. Moreover, time complexity of the PSM algorithm is much less compare to the KNN due to search of selective genes for imputing missing value.

REFERENCES

- [1] D.J. Lockhart and E. A. Winzeler, "Genomics, gene expression and DNA arrays," Nature, vol. 405, pp. 827-836, 2000.
- [2] M. Ouyang, A.T. Garnett, T.M. Han, K. Hama, A. Lee, Y. Deng, N. Lee, H.Y. Liu, S.L. Amacher, S.A. Farber, and S.Y. Ho, "A web based resource characterizing the zebrafish developmental profile of over 16,000 transcripts," Gene Expr. Patterns, vol. 8(3), pp. 171-180, 2008.
- [3] A.G. Brevern, S. Hazout and A. Malpertuy, "Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering," BMC Bioinformatics, vol. 22(5), pp. 566-572, 2004.
- [4] E. Wit and J. McClure, "Statistics for Microarrays Design, Analysis and Inference," John Wiley & Sons, 2004.
- [5] A. J. Butte, and J. Ye, "Determining significant fold differences in gene expression analysis," Pac. Symp. Biocomput., vol. 6, pp. 6-17, 2001.
- [6] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown and L. M. Staudt, "Distinct types of diffuse large B-cell

- lymphoma identified by gene expression profiling," Nature, vol. 403, pp. 503-511, 2000.
- [7] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art", Psychol. Methods, vol. 7, pp. 144-177, 2002.
- [8] P. Meesad, K. Hengpraprom, "Combination of KNN-Based Feature Selection and KNNBased Missing-Value Imputation of Microarray Data," Innovative Computing Information and Control, ICICIC apos, vol. 18(20), pp. 341-341, 2008.
- [9] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. B. Altman, "Missing value estimation methods for DNA microarrays," BIOINFORMATICS, vol. 17, pp. 520-525, 2001.
- [10] J. Luo, T. Yang and Y. Wang, "Missing Value Estimation For Microarray Data Based On Fuzzy C-means Clustering," The Eighth International Conference on High-Performance Computing in Asia-Pacific Region, 2005.
- [11] S. Zhang, J. Zhang, X. Zhu, Y. Qin and C. Zhang, "Missing Value Imputation Based on Data Clustering," Transactions on Computational Science I, vol. 4750, pp. 128-138, Springer, Heidelberg, 2008.
- [12] J. Liu and W. Wang, "Op-Cluster: Clustering by Tendency in High-Dimensional Space," Proc. Third IEEE Int'l Conf. Data Mining (ICDM '03), pp. 187-194, 2003.
- [13] J. Liu, J. Yang and W. Wang, "Biclustering in Gene Expression Data by Tendency," Proc. Third IEEE Int'l Computational Systems Bioinformatics Conf. (CSB '04), pp. 182-193, 2004.
- [14] H. Wang, W. Wang, J. Yang, and P.S. Yu, "Clustering by Pattern Similarity in Large Data Sets," Proc. ACM SIGMOD '02, pp. 394-405, 2002.
- [15] S. Erdal, O. Ozturk, D. Armbruster, H. Ferhatosmanoglu, W.C. Ray, "A Time Series Analysis of Microarray Data," Proc. Fourth IEEE Int'l Symp. Bioinformatics and Bioeng. (BIBE '04), pp. 366-378, 2004.
- [16] J. Pei, X. Zhang, M. Cho, H. Wang, P.S. Yu, "Maple: A Fast Algorithm for Maximal Pattern-Based Clustering," Proc. Third IEEE Int'l Conf. Data Mining (ICDM '03), pp. 259-266, 2003.
- [17] H. Wang, F. Chu, W. Fan, P.S. Yu and J. Pei, "A Fast Algorithm for Subspace Clustering by Pattern Similarity," Proc. 16th Int'l Conf. Scientific and Statistical Database Management (SSDBM '04), pp. 51-62, 2004.
- [18] L. Zhao and M.J. Zaki, "Tricuster: An Effective Algorithm for Mining Coherent Clusters in 3D Microarray Data," Proc. ACM SIGMOD '05, pp. 51-62, 2005.
- [19] Y. Zhang, H. Zha and C.H. Chu, "A Time-Series Biclustering Algorithm for Revealing Co-Regulated Genes," Proc. IEEE Int'l Symp. Information Technology: Coding and Computing (ITCC '05), pp. 32-37, 2005.
- [20] H. Yu, N.M. Luscombe, J. Qian and M. Gerstein, "Genomic Analysis of Gene Expression Relationships in Transcriptional Regulatory Networks," Trends Genet, vol. 19(8), pp. 422-427, 2003.
- [21] http://biodata.mshri.on.ca/yeast_grid/servlet/
- [22] D. Dembélé and P. Kastner, "Fuzzy c-means method for clustering microarray data," Bio-Informatics, vol. 19(8), PP 973-980, 2003.
- [23] IEEE Computer Society : IEEE Standard for Binary Floating-Point Arithmetic, IEEE Std 754-2008, 2008.
- [24] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Amders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," Mol Biol Cell, vol. 9, pp. 3273-3297, 1997.