IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

Handling Missing Values based on K-NN Imputation with Regression

R. Kavitha[1], S. Sandhya[2]

Assistant Professor, Dept. of Computer Science, Sri Krishna Arts & Science College, Coimbatore, India[1]

M. Phil Scholar, Dept. of Computer Science, Sri Krishna Arts & Science College, Coimbatore, India[2]

Abstract: The present information era needs knowledge discovery from the vast volume of data. As computer technology has developed to greater height, specifically the Internet led to bang of data. Data availability has gone beyond the human capability of absorption. This increase in enormous volume and varied data paves the way for advances in method to recognize, develop and summarize the data. The data set taken from the microarray experiments often contain some missing values which may primarily occur due to scratches or spots on the slide, dust, inadequate resolution, image corruption and hybridization failures. In this paper, a novel approach is proposed for estimating (predicting) missing values using k-NN regression imputation method to handle incomplete data set. The proposed work provides considerably better results when compared to existing work.

Keywords: k-NN with regression, missing value, data mining, microarray.

I. INTRODUCTION

It is common to find missing values available in microarray data due to scratches or spots on the slide, or image corruption that occur during experiments [1]. From an estimate on microarray datasets it is noted that 5% to 80% of genes have one or more missing values [2] [3]. Hence microarray data desires to be preprocessed in handling or predicting the missing value. The first method to handle the missing value is by eliminating the records which have even a single missing value they construct invalid results [4]. The second approach involves a numerical substitution by a constant 0 [4] or mean substitution [5], which may misinterpret the association among variables. Imputation is the third approach that selects the gene with missing value first and predicts them using the observed values of selected gene.

There are cluster-based algorithms [6] [7] that deal with missing values with no user-specific parameters. In this paper, a novel approach is proposed for estimating (predicting) missing values using k-NN regression imputation method to handle incomplete data set.

II. LITERATURE REVIEW

The multivariate analysis methods generate undesirable results due to the high percentage of missing values, e.g. hierarchical clustering and the support vector machine classifier [8] [9]. Moreover many analysis methods such as principal component analysis (PCA), singular value decomposition (SVD) and generalized SVD (GSVD) have given invalid results when applied to datasets with missing values [10] [11].

Imputation method improves the estimation by incorporating the correlation structure between entries in the gene expression data matrix and domain knowledge about data. Several imputation methods are proposed such as weighted k-nearest neighbors (kNNimpute) (Troyanskaya et al 2001), singular value decomposition (SVD) [12], least squares imputation(LSimpute) [13], Bayesian principal component analysis(BPCA) [14], Gaussian mixture clustering imputation(GMCimpute) [3], collateral missing value imputation(CMVE) [15] and weighted nearest neighbors imputation (WeNNI)[16].

Even though k-NN is an effective method which finds k nearest instances the training sample for classifying unknown data instance, it relies on sub-groups for decision of unlabelled samples. To overcome this, a computationally intensive approach that asymptotically improves the performance of nearest neighbor classifiers with k-NN regression imputation method is proposed.

III.METHODOLOGY

The proposed work undergoes data collection, transformation, missing value handling and prediction models using clustering techniques. The data set is collected from the public data set of Gene Expression Pattern Analysis Suite –

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

Yeast cell cycle(http://transcriptome.ens.fr/gepas/docs/preprocess/index.html, http://kzi.polsl.pl/~jbiesiada/Infosel/files/datasets.html).
The collected raw dataset is transformed using min-max normalization and discretized. Then the missing value is treated with the existing and proposed methods and the results are analyzed with cluster distance and RMSE measures.

a. Data Transformation

Data transformation undergoes normalization and discretization. Normalization is a data preprocessing tool used in data mining system. The normalization process undergoes transforming the values of the variable in the dataset such that they lie within the specified range of 0 to 1. The techniques such as nearest neighbor classification and clustering use normalization process. Min-max normalization, z-score normalization and normalization by decimal scaling are commonly used data normalization methods. The min-max normalization and equal frequency discretization used in this proposed algorithm are given in the next section.

b. Min-Max Normalization

Min-max normalization maps a value x of variable v to x′ in the range [new_min(v), new_max(v)]. The min-max normalization is calculated by the following formula:

$$x' = \frac{[x - \min(v)]}{[\max(v) - \min(v)]}[new\_\max(v) - new\_\min(v)] + new\_\min(v) \quad (3.1)$$

where min(v) = minimum value of variable, max(v) = maximum value of variable.

c. Discretization

The purpose of discretization process is to induce a list of intervals that divide the numerical domain of a continuous explanatory attribute. The equal-frequency discretization algorithm calculates the minimum and maximum values of the discretized attribute, sorts all values in ascending order, and divides the sorted values into k intervals so that every interval contains approximately the same number of values. Thus each interval contains n=k (possibly duplicated) adjacent values where k is a user predefined parameter.

d. Prediction Models Using Clustering Technique

Data clustering aims at grouping of objects such that data points belonging to one cluster are similar whereas objects in different clusters are distinct. Clustering can define dense and sparse regions to identify overall distribution patterns and interesting correlations among data attributes. The clusters are created to optimize the similarity function such as distance so that objects within the cluster are similar. The selection of clustering algorithm depends on type of the data, purpose and application. The different clustering techniques are partitioning, hierarchical, density-based methods, grid-based methods and model-based methods.

- Random Clustering is a simple clustering process that creates simple and uniform random partitions. A random data set is given as input to this process. The desired number of clusters is assigned as an input parameter value. A random seed is specified for the clustering process.
- Density Based Spatial Clustering of Applications with Noise is a density-based method which aims at discovering arbitrary shape clusters [17]. The clusters are represented as dense region of objects in the object space separated from low density regions. It defines clusters as density-connected points. A point which is not contained in any of the cluster is considered to be noise. It doesn't require the number of clusters to be known in prior unlike in k-means algorithm [18]. Euclidean distance measure is most commonly used distance metric by DBSCAN.
- Support vector clustering (SVC) does not take user specific parameters such as number or shape of clusters, which is more suitable for low-dimensional data. Kernel function is used to map data space to high dimensional feature space. SVC uses Support Vector Domain Description (SVDD) to delineate the region in data space where the input examples are concentrated. It belongs to the general category of kernel based learning.
- Farthest first is a variant of k-means which points each cluster centre in turn at the point furthermost from the existing cluster centre. This point must lie within the data area. This greatly speeds up the clustering in most of the cases since it undergoes less reassignment.

e. k-NN impute algorithm

The k-NN-based method selects incomplete genes to impute missing values with expression profiles similar to the gene of interest. To determine the missing value of a specific gene, the k-NN impute searches from the pattern space of training samples the most closest to the unknown gene. The "closeness" is defined by the Euclidean distance d(A,B) which is given as

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

$$d(A,B)= \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \qquad (3.2)$$

where A=(a1,a2,…an) and B=(b1,b2,…bn). The algorithm returns the average value of the prediction in association with the k-nearest neighbors of the missing gene.

f. Proposed Approach – Regression with k-NN Imputation(RkNN)

This section discusses on the proposed approach of enhancing k-NN approach with the help of regression method.
In the proposed method regression on every pair of variables is used as a learner. A sample is generated by uniformly sampling m gene expressions from the training set with replacement and T instances RS1, RS2,…,RST are generated and a classifier Ci is built from each sample Ri, i=1,2,…T. A final classifier C* is built from C1, C2… CT whose output is the improved composite classifier.

Algorithm 3.1 Regression with k-NN imputation
1. Create 't' data sets from a database applying the sampling with replacement scheme
2. Apply k-NN impute to each sample training data set to generate classifiers.
Classification using k-NN imputation is as follows:
- Compute Euclidean distance between X and all the training expressions
- Sort the distance and determine k nearest neighbors based on maximum distance
- Assign the most similar complete gene expression among its k nearest neighbors

3. Choose instances RSi (i=1,2,…t) of size n from the original training set S
4. Compute average of each classifier prediction to obtain composite classifier.
5. For an object with unknown decision, make predictions with each of the t classifiers.
6. Select the most frequently predicted decision based on mean of the k-most similar instances.

IV. EXPERIMENTAL STUDIES

The experiment is conducted on the publicly available yeast cell cycle dataset which contains 6178 genes with 77 samples. Colon cancer dataset contains 62 samples with 2000 genes. Leukemia dataset consists of 38 samples with 7129 genes. Splice dataset contains 3190 samples with 60 genes. The four gene microarray datasets are varied with 5% to 25% missing values for analysis.

From the microarray datasets, samples are generated with replacement scheme. Four learners were generated from yeast cell cycle, colon cancer and leukemia datasets respectively. Fifty learners were generated for splice dataset. k-NN imputation is used for building all the learners. 60% of sample data is assigned for training and remaining 40% is taken for testing.

The performance metrics used in this experiment are cluster distance and RMSE. The minimum cluster distance shows the closeness of objects within the cluster. In this research work the implementation of algorithms are done by a machine learning algorithm tool called Rapid Miner version 5.

Table 1 Mean cluster distance on yeast cell cycle data set

Clustering techniques	Mean Substitution	k-NN Imputation	Enhance k-NN with regression
Random Clustering	200.63	195.72	107.82
Support Vector Clustering	293.64	208.71	151.48
DBSCAN	369.34	342.43	214.38
Farthest first	535.62	443.68	80.82

Table 2 Mean cluster distance on Leukemia data set

Clustering techniques	Mean Substitution	k-NN Imputation	Enhance k-NN with regression
Random Clustering	319.58	272.49	168.22
Support Vector Clustering	468.28	387.47	106.38
DBSCAN	583.44	463.88	349.40
Farthest first	552.68	459.42	204.65

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

Table 1 provides the results of the performance of mean substitution method, k-NN imputation method and enhanced k-NN with regression method for average cluster distance over yeast cell cycle dataset. The results showed that enhanced k-NN with regression gives considerably better results compared with the k-NN imputation method.

Table 2 shows the average cluster distance over colon cancer dataset for mean substitution method, k-NN imputation method and enhanced k-NN with regression method. The results revealed that enhanced k-NN with regression provides significantly better when compared with the other existing methods.

The average cluster distance on Leukemia data set is shown in Table 3 using mean substitution method, k-NN imputation method and enhanced k-NN with regression method. From the results it is inferred that enhanced k-NN with regression method provides significantly better results when compared with the other existing methods.

Table 3 Mean cluster distance on colon cancer data set

Clustering techniques	Mean Substitution	k-NN Imputation	Enhance k-NN with regression
Random Clustering	655.29	401.39	204.51
Support Vector Clustering	535.14	332.42	190.31
DBSCAN	453.83	423.21	228.78
Farthest first	390.38	359.20	165.32

Table 4 provides the results of the average cluster distance on splice data set using mean substitution method, k-NN imputation method and enhanced k-NN with regression method. The results reveal that enhanced k-NN with regression performs better when compared with the other existing methods.

Table 4 Mean cluster distance on splice data set

Clustering techniques	Mean Substitution	k-NN Imputation	Enhance k-NN with regression
Random Clustering	537.42	455.93	150.64
Support Vector Clustering	839.89	530.51	263.62
DBSCAN	672.62	592.46	249.26
Farthest first	471.34	432.42	186.32

Table 5 Evaluation results on yeast cell cycle data set

Evaluation Criteria	Mean Substitution	SVD	k-NN Imputation	Regression k-NN imputation
NRMSE	0.43	0.57	0.43	0.25
UCE	0.66	0.63	0.59	0.34
SCE	0.94	0.63	0.57	0.28
RMSE	0.55	0.79	0.43	0.15

Table 6 Evaluation results on Colon cancer data set

Evaluation Criteria	Mean Substitution	SVD	k-NN Imputation	Regression k-NN imputation
NRMSE	0.27	0.37	0.20	0.10
UCE	0.58	0.59	0.48	0.22
SCE	0.66	0.73	0.68	0.38
RMSE	0.45	0.64	0.51	0.25

Table 7 Evaluation results on Leukemia data set

Evaluation Criteria	Mean Substitution	SVD	k-NN Imputation	Regression k-NN imputation
NRMSE	0.85	0.68	0.69	0.32
UCE	0.93	0.55	0.40	0.18
SCE	0.72	0.52	0.47	0.26
RMSE	0.88	0.63	0.58	0.25

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

Table 8 Evaluation results on Splice data set

Evaluation Criteria	Mean Substitution	SVD	k-NN Imputation	Regression k-NN imputation
NRMSE	0.83	0.73	0.54	0.29
UCE	0.70	0.78	0.60	0.48
SCE	0.81	0.65	0.51	0.24
RMSE	0.92	0.87	0.74	0.35

The performance of the missing value estimation is evaluated by the normalized root mean squared error (NRMSE):

$$NRMSE = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}(R_i - I_i)^2}{\text{var}\,iance[R_i]}}$$

where Ri is the real value and Ii is the imputed value. The mean and the variance are computed from the missing entries in the data set. The imputation method attaining the smallest RMSE gives the most correct prediction of the complete data set when estimated values are included.

Figure 1-4 shows the performance of the four missing value handling methods on yeast cell cycle, colon cancer, leukemia and splice data set respectively. The horizontal and vertical axes indicate the percentage of missing entries in the data matrix and the NRMSE of estimation methods respectively.

Unsupervised classification error(UCE) evaluates the preservation of the internal construction by measuring how good the clustering of the complete dataset was well-preserved when clustering the imputed dataset.

UCE = percentage of misclassified instances

Supervised classification error(SCE) evaluates the preserving of discriminative or prediction capability by measuring the variation between subclasses predicted by supervised classification after missing data imputation and the predicted subgroups.

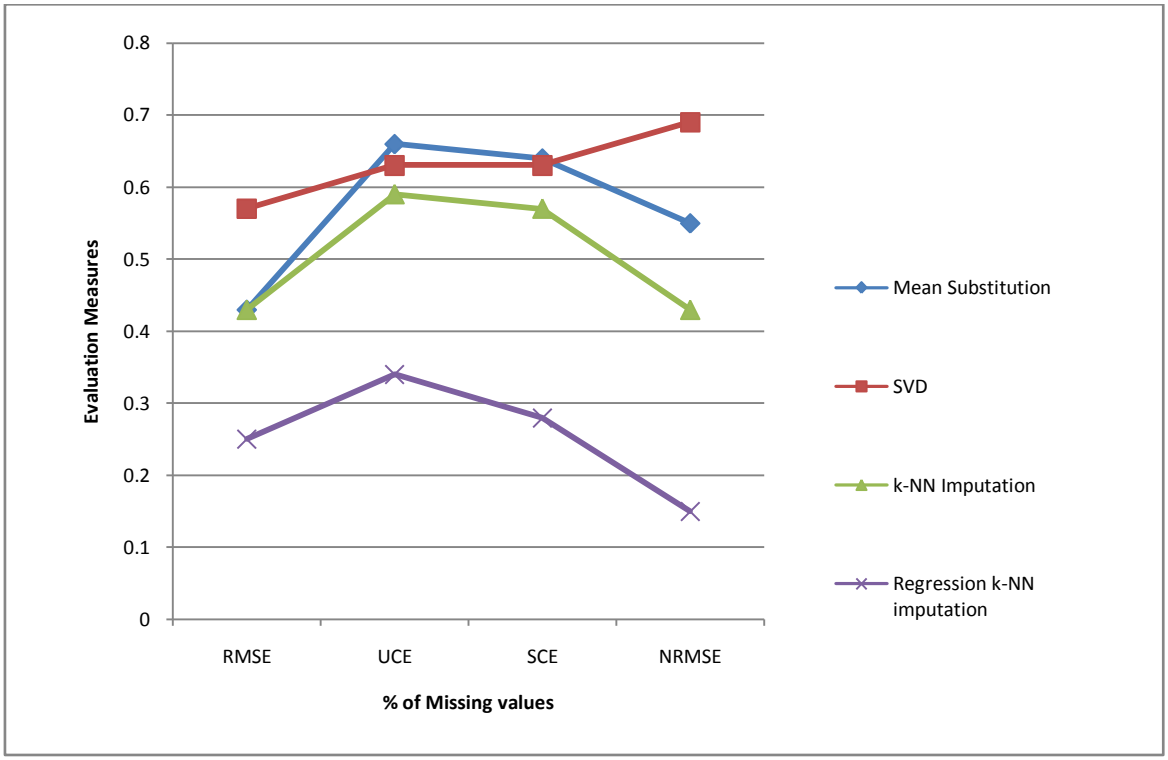

Figure 1 Evaluation of missing values using RMSE, UCE, SCE, NRMSE for yeast cell cycle data

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

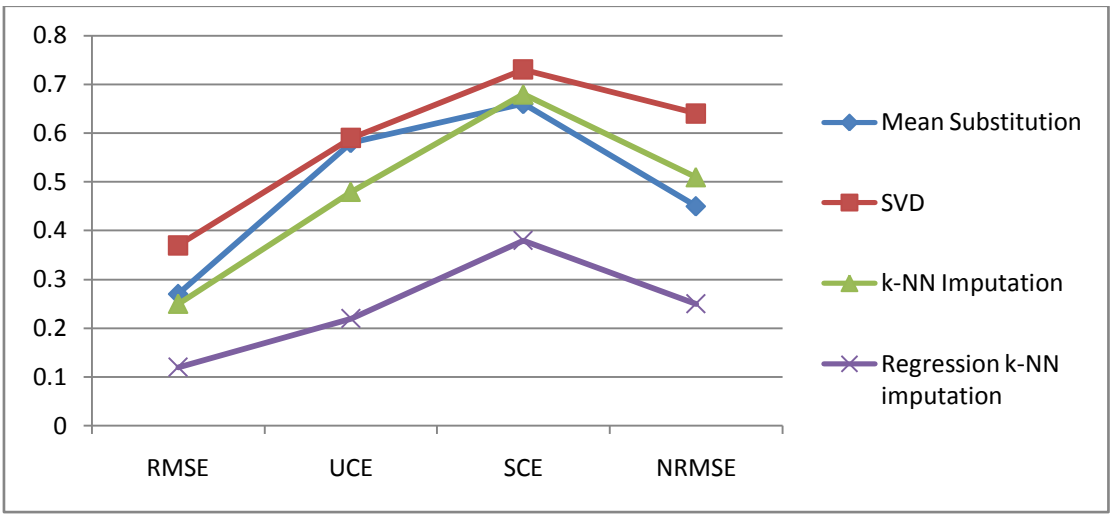

Figure 2 Evaluation of missing values using RMSE, UCE, SCE, NRMSE for colon cancer data.

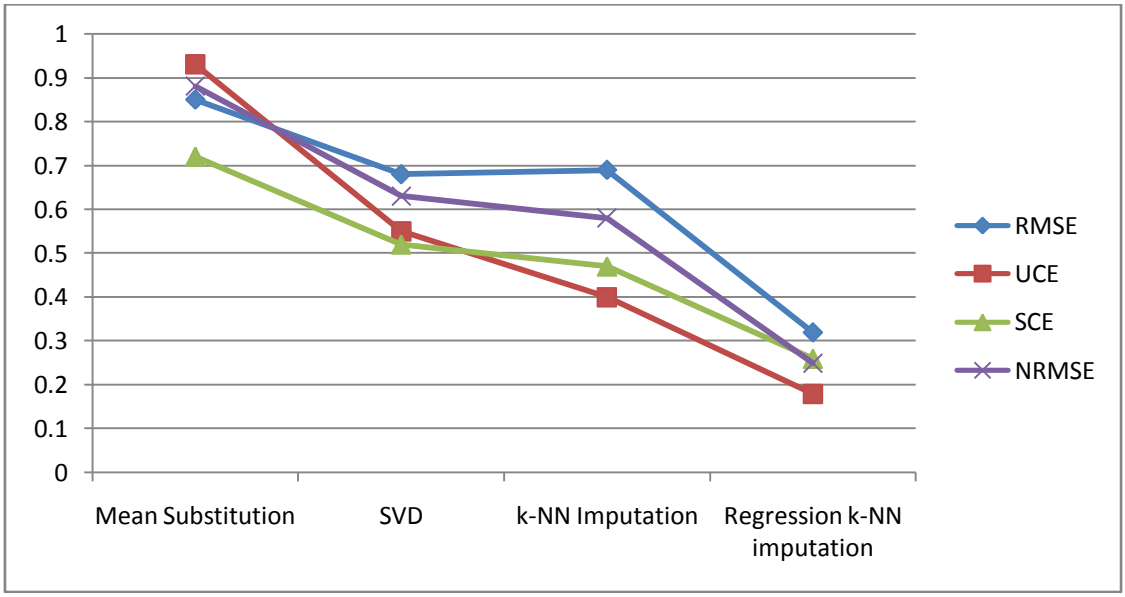

Figure 3 Evaluation of missing values using RMSE, UCE, SCE, NRMSE for leukemia data

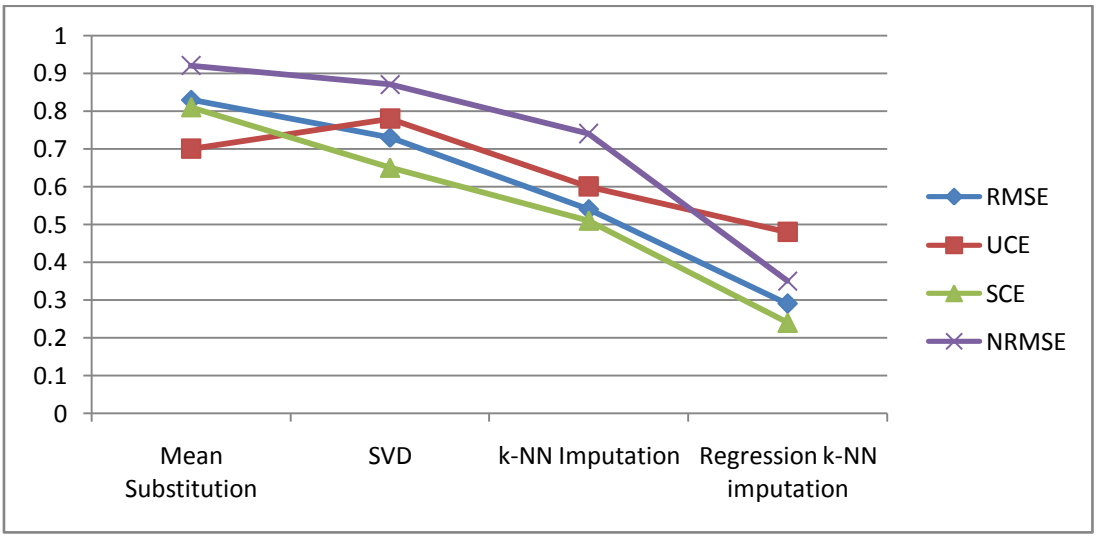

Figure 4 Evaluation of missing values using RMSE, UCE, SCE, NRMSE for splice data

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

V. CONCLUSION

In this paper missing value handling is explored using the different techniques. Gene expression data suffer from missing values which may affect subsequent analysis. As k-NN is an unstable learning algorithm and regression works well for "unstable" learning algorithms, an enhanced regression with k-NN imputation method is proposed to predict the missing values. So instead of using k-NN imputation the performance of predicting missing value is considerably improved by regression with k-NN imputation method. The performance evaluation using cluster distance and NRMSE show that the proposed method provides significantly better results when compared with the existing methods.

REFERENCES

[1] Kim, KY, Kim, BJ & Yi, GS 2004, 'Reuse of imputed data in microarray analysis increases imputation efficiency', BMC Bioinformatics, vol. 5, no. 160.
[2] Tuikkala, J, Elo, L, Nevalainen, OS & Aittokallio, T 2006, 'Improving missing value estimation in microarray data with gene ontology', Bioinformatics, vol. 22, no. 5, pp. 566-572.
[3[Ouyang, M, Welsh, WJ & Georgopoulos, P 2004, 'Gaussian mixture clustering and imputation of microarray data', Bioinformatics, vol. 20, no. 6, pp. 917-923.
[4] Alizadeh, AA, Eisen, MB, Davis, RE, Ma, C, Lossos, IS, Rosenwald, A, Boldrick, JG, Sabet, H, Tran, T, Yu, X, Powell, JI, Yang, LM, Marti, GE, Moore, T, Hudson, J, Lu, LS, Lewis, DB, Tibshirani, R, Sherlock, G, Chan, WC, Greiner, TC, Weisenburger, DD, Armitage, JO, Warnke, R, Levy, R, Wilson, W, Grever, MR, Byrd, JC, Botstein, D, Brown, PO & Staudt, LM 2000, 'Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling', Nature, vol. 403, no. 6769, pp. 503-611.
[5] Schafer, JL & Graham, JW 2002, 'Missing data: Our View Of The State of the Art', Psychol. Methods, vol.,7, pp. 144- 177.
[6] Luo, J, Yang, T & Wang, Y 2005, 'Missing Value Estimation For Microarray Data Based on Fuzzy C-means Clustering', Proceedings of the Eighth International Conference on High-Performance Computing in Asia-Pacific Region.
[7] Zhang, S, Zhang, J, Zhu, X, Qin, Y & Zhang, C 2008, 'Missing Value Imputation Based on Data Clustering, Transactions on Computational Science (TCOS), vol. 1, pp. 128-138.
[8] Brown, MPS, Grundy, WN, Lin, D, Cristianini, N, Sugnet, CW, Furey, TS, Ares, M & Haussler, D 2000, 'Knowledge-based analysis of microarray gene expression data by using support vector machines', proceeding of National Acad. Sci., vol. 97, no. 1, pp. 262-267.
[9] Eisen, MB, Spellman, PT, Brown, PO & Botstein, D 1999, 'Cluster analysis and display of genome-wide expression patterns', proceedings of NatlAcadSci, vol. 95, no. 25, pp. 10863-10868.
[10] Raychaudhuri, S, Stuart, JM & Altman, RB 2000, 'Principal components analysis to summarize microarray experiments: application to sporulation time series', Pac SympBiocomput, pp. 455-466.
[11] Alter, O, Brown, PO & Botstein, D 2003, 'Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms', proceedings of the national academy science, vol. 100, no. 6, pp. 3351-3356.
[12] Troyanskaya, O, Cantor, M, Sherlock, G, Brown, P, Hastie, T, Tibshirani, R, Botstein, D & Altman, RB 2001, 'Missing Value Estimation Methods for DNA Microarrays', Bioinformatics, vol. 17, pp. 520-525.
[13] Bo, T, Dysvik, B & Jonassen, I 2004, 'LSimpute: accurate estimation of missing values in microarray data with least squares methods', Nucleic Acids Research, vol. 32.
[14] Oba, S, Sato, M, Takemasa, I, Monden, M, Matsubara, K & Ishii, S 2003, 'A Bayesian missing value estimation method for gene expression profile data', Bioinformatics, vol. 19, no. 16, pp. 2088-2096.
[15] Sehgal, MS, Gondal, I & Dooley, L 2005, 'Collateral missing value estimation: Robust missing value estimation for consequent microarray data processing', Advances In Artificial Intelligence, vol. 3809, pp. 274-283.
[16] Johansson, P & Hakkinen, J 2006, 'Improving missing value imputation of microarray data by using spot quality weights', BMC Bioinformatics, vol. 7.
[17] Ester, M, Kriegel, HP, Sander, J & Xu, X 1996, 'A density-based algorithm for discovering clusters in large spatial databases with noise', Data Mining and Knowledge Discovery, pp. 226-231.
[18] MacQueen, JB 1967, 'Some Methods for classification and Analysis of Multivariate Observations', Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press. pp. 281–297.