The Korean Journal of Applied Statistics (2011)
24*(2), 315–321*
DOI: 10.5351/KJAS.2011.24.2.315

Verification of Improving a Clustering Algorithm for Microarray Data with Missing Values

SuYoung Kim[1]

[1]Center for Korean Studies Materials, The Academy of Korean Studies

(Received November 2010; accepted January 2011)

Abstract

Gene expression microarray data often include multiple missing values. Most gene expression analysis (including gene clustering analysis); however, require a complete data matric as an input. In ordinary clustering methods, just a single missing value makes one abandon the whole data of a gene even if the rest of data for that gene was intact. The quality of analysis may decrease seriously as the missing rate is increased. In the opposite aspect, the imputation of missing value may result in an artifact that reduces the reliability of the analysis. To clarify this contradiction in microarray clustering analysis, this paper compared the accuracy of clustering with and without imputation over several microarray data having different missing rates. This paper also tested the clustering efficiency of several imputation methods including our proposed algorithm. The results showed it is worthwhile to check the clustering result in this alternative way without any imputed data for the imperfect microarray data.

Keywords: Microarray, gene expression, clustering, missing value.

1. Introduction

Clustering is one of the core analyses for DNA microarray data. Clustering of microarray data is used to find groups of similarly expressed genes among thousands of genes that might elucidate the functional relationships within or among the groups of genes. The data from microarray experiments is usually in the form of large matrices of expression levels of genes (rows) under different samples (columns) and frequently with some values missing. To measure the similarity of genes, most of the clustering methods use correlation which cannot be computed for the genes having missing values. Missing value occurs commonly in microarray data and sometimes it is serious as more than 90% of genes have missing values (Ouyang *et al.*, 2004).

A typical solution to handle missing entries in a dataset is imputation. The missing value can be predicted reasonably based on the values of the other genes in the same group. Many different types of imputation methods have been introduced in microarray data analysis (Troyanskaya *et al.*, 2001). However, imputation has an intrinsic limitation that could mislead the clustering result due to the use of estimated values. This becomes worse when the missing rate of the dataset increases

[1]Researcher, Culture Informatics Division, The Academy of Korean Studies, 110 Haogogae-gill, Bundang-gu, Seongnam-si, Gyeonggi-do 463-791, Republic of Korea. E-mail: tmt21@aks.ac.kr

or the missing values are severely localized on the one side of data matrix. An adverse effect of imputation for the clustering of microarray data should be suspected in these circumstances. In addition, the imputation needs almost the same amount of computation as the clustering method itself.

Until now, the alternative approach which can cluster the microarray data directly without imputation has not been seriously studied. It is required to verify the effect of a possible alternative approach as well as the imputation on the clustering of imperfect microarray data.

2. Research Questions

We have developed a new gene clustering method based a one dimensional sample, which directly cluster the microarray data having missing entries without imputation process. A gene with any missing entry should be removed or imputed because the complete elements are necessary to compute the distance of the multi-dimensional vector. However, our method could handle the missing entries by the decomposition of multi dimensional data (genes for multiple samples) into one dimensional data (genes for a sample). Initial clusters are generated for each sample without imputing the missing values. Because each gene has multiple samples, the ignored gene in one sample by the missing value can contribute to the clustering with the values in the other samples. In this way, the method could fully use the remaining information even if a gene has missing values. The proposed method(PM) can use any conventional clustering method to generate the initial clusters for each sample. After the generation of the initial clusters for each sample, the cluster membership of each gene is determined by combining the information of the initial clusters. The final assignment of the cluster of a gene is decided based on the frequency and the validity of the assigned clusters for the gene in the individual samples. The overall computational complexity is almost the same as the typical clustering method.

There are clustering methods that include the imputation of the missing value inside of the clustering process (Kim *et al.*, 2006). This type of method tries to reduce the bad effect of improper imputation by the iterative estimation of imputed values and optimized clusters. However, the computational complexity is very high and they still include imputed values.

The validity of the proposed method(PM) which does not utilize any estimated value for the missing entries during clustering process was tested with a model data and several real microarray data. The effect of the various missing rate on the performance of different clustering methods was investigated as well.

3. Research Algorithm and Validation

In the proposed method, each sample (column) of microarray data was initially clustered with existing clustering algorithm and the validity of clusters of each sample was evaluated by gap statistic (Tibshirani *et al.*, 2001). The final cluster membership for a gene was decided by polling the indices of clusters built in individual samples weighted by the validity of containing clusters and the completeness of entries. The detailed process is as follows.

(1) Let $E = \{E_{ij},\, i = 1, \ldots, N,\, j = 1, \ldots, M\}$ be a gene expression matrix with gene set $G = \{g_1, \ldots, g_i, \ldots, g_N\}$ and sample set $S = \{S_1, \ldots, S_j, \ldots, S_M\}$. At first, the expression values $E_i = \{E_{1j}, \ldots, E_{ij}, \ldots, E_{Nj}\}$ of sample j is clustered into K clusters $X_k = \{X_{j1}, X_{j2}, X_{j3}, \ldots, X_{jK}\}$ with a given clustering method. In our test, we used the K-means (Hartigan and Wong,

1979) algorithm for the clustering.

(2) Then, the Gap statistics, GP_{jk} for each cluster X_{jk}, $k = 1, 2, 3, \ldots, K$ in each sample are calculated. The optimal number (k') and the validity ($\mathrm{GP}_{jk'}$) of the given cluster set for sample j are determined by gap statistic such that

$$k' = \text{the smallest } k \text{ such that } \mathrm{GP}_{jk} \geq \mathrm{GP}_{jk+1} - \mathrm{sd}_{k+1},$$

where sd is the standard deviation of within dispersion measure of the cluster. The number of clusters was tested up to 50 ($K = 50$) clusters for each sample.

(3) The optimal cluster number of whole samples is determined by the voting of samples weighted by the gap statistic score. The optimal number ($k_{j'}$) of clusters for individual sample is weighted by $f_{kj'} = \mathrm{GP}_{jk'}/(\Sigma_j \mathrm{GP}_{jk'})$. The optimal cluster number (K) of the whole sample is chosen via,

$$K = \mathrm{argmax}\{k_{j'} | \Sigma_{kj'} f_{kj'}\}$$

means the cluster number having the highest frequency and gap statistic ratio.

(4) The sample that has the lowest number of missing entries among the samples having K (the optimal number) clusters with highest $f_{Kj'}$ becomes the representative sample, R. The cluster index ($C_{jh} = 1, \ldots, K$) of sample j is assigned to the corresponding cluster index ($C_k = C_1, \ldots, C_K$) of the representative sample R in the condition that the intersection of matched members is maximized.

(5) The final cluster membership (C_{gi}) for a gene, g_i is decided by the weighted polling of the cluster index (C_k) assigned for an individual sample j such that

$$C_{gi} = \mathrm{argmax}\{C_k | \Sigma_{C_k} f_{kj'}\}.$$

The proposed clustering method was applied to five different datasets contained one supervised data and four unsupervised data. The well known iris dataset was selected as a model dataset for the primary test of the performance of methods.

This dataset is obtained from UCI Repository Of Machine Learning Databases and Domain Theories (ics.uci.edu: pub/machine-learning-databases). The dataset, which is supervised data, contains three clusters of iris species setosa, versicolor, and virginica with 50 instances each which are well separated by four different features of flowers. To evaluate the methods with real world microarray data, we chose four different types of microarray datasets. SRBCT dataset has 2308 genes and 63 experimental conditions with 8 Burkitt Lymphoma(BL), 23 Ewing Sarcoma(EWS), 12 neuroblastoma(NB), and 20 rhabdomyosarcoma(RMS) samples (Khan *et al.*, 2001). Colon dataset consists of 2000 genes using an Affymetrix Oligonucleotide array from 22 normal and 40 colon tumor tissues (Alon *et al.*, 1999). Five tissues dataset has 3529 genes and 10 samples; 2 samples for each of 5 subclasses (testes, brain, liver, muscle and bone marrow) (Le *et al.*, 2004). The breast cancer dataset has 3226 genes and 22 samples (Hedenfalk *et al.*, 2001). Table 3.1 shows summary of five datasets in this paper.

The basic goal of this research was to verify if the proposed method(PM) could alternate the conventional clustering methods that need imputation. At first, the innate clustering performance of the proposed method was examined with a supervised and complete dataset. Because the clusters of the iris data is already known, we could estimate the clustering accuracy of the methods by

Table 3.1. Summary of the dataset used in this study

Dataset	Size(indivisduals $*$ samples)	Classes (♯ of samples)
Iris	$150 * 4$ Supervised data	Sepal length (1) Sepal width (1) Petal length (1) Petal width (1)
Small Round Blue Cell Tumo (SRBCT)	$2308 * 63$ Unsupervised data	Burkitt lymphoma (8) Neuroblastoma (12) Rhabdomyosarcoma (20) Ewing sarcoma (23)
Colon	$2000 * 62$ Unsupervised data	Normal (22) Tumor colon tissues (40)
Breast Cancer	$3226 * 22$ Unsupervised data	BRACA1 (7) BRACA2 (8) Sporadic (7)
Five Tissues	$3529 * 10$ Unsupervised data	Testes (2) Brain (2) Liver (2) Muscle (2) Bone marrow (2)

Table 3.2. The performances of various clustering methods on iris datasets.

Method	Average silhouette coefficient	Accuracy (%)
PM	0.55	96
K-means	0.55	91.33
SVD	0.54	90.7
Hierarchical	0.51	89.4
Diana	0.53	92.1
Fuzzy	0.54	92

comparing with the true clusters (three clusters of iris species) suggested from the dataset itself. The validity of clusters was evaluated by a silhouette coefficient that measures the quality of a clustering result by the within cluster compactness and the inter cluster separation (Rousseeuw, 1987). The average silhouette coefficient of the resulting clusters was used to validate each method. Table 3.2 shows that the validity of clusters and the accuracy of the proposed method outperforms against the other methods such as K-means, Singular Value Decomposition based method(SVDimpute), Hierarchical, DIANA(DIvisive ANAlysis), and Fuzzy clustering methods. In these methods, SVD imputation attempts to utilize the global information in the entire matrix in predicting the missing values. The basic concept is to find the dominant components summarizing the entire matrix and then to predict the missing values in the target genes by regressing against the dominant components. If we perform SVD to matrix Y, we get the following equation.

$$A_{M\times N} = U_{M\times M}\Sigma_{M\times N}V^T_{N\times N},\ \text{(Alter } \textit{et al.}\text{, 2000; Anderson, 1984; Golub and Van Loan, 1996).}$$

Let $L = \min\{M, N\}$, matrix V^T now contains L eigengenes v_l $(0 < l < L)$, and matrix U contains L eigenarrays u_l $(0 < l < L)$. In the algorithm SVDimpute, the k most significant eigengenes from V^T are selected, and missing value $A_{i,j}$ is estimated by first regressing the expression profile vector of gene i against the K eigengenes and then using the coefficients of the regression to reconstruct $\hat{A}_{i,j}$ from a linear combination of the K eigengenes. If we denote the expression profile vector of gene i in A as a and assume that v_l $(l = 1, 2, \ldots, K)$ are the eigengenes, $\tilde{v}_l$ and $\tilde{a}$ are vectors that are

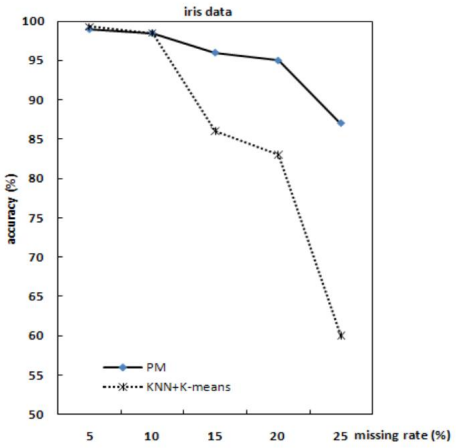

Figure 3.1. The accuracy measure for proposed method(PM) and K-means method combined with KNN imputation (KNN+k-means) on iris dataset with different missing rates.

obtained by deleting the j^{th} component of v_l and a, then the missing component $A_{i,j}$ is estimated as follows.

$$\hat{A}_{i,j} = \sum_{l=1}^{k} \left(\tilde{v_l}^T \bullet \tilde{a} \right) v_{l,j}, \text{ (Gan } et \text{ } al., \text{ 2006).}$$

Since SVD can only be performed on complete matrices, the iterative expectation maximization method is used. The hierarchical clustering algorithm used is based on the average linkage method.

All of these methods were selected because they have been introduced in microarray data analysis before (Kaufman and Rousseeuw, 1990). Interestingly, the silhouette coefficient of K-means that was used in our clustering step shows a tie with ours. The accuracy improvement was prominent against conventional clustering methods. The results verify that the method combining one-dimensional clusters which has never been introduced in microarray data analysis is not inferior to the other conventional methods that use the multi-dimensional sample space together. The advantage of our algorithm was significant when it was applied to the incomplete datasets with missing entries. Figure 3.1 shows that the accuracy of our algorithm and K-means clustering combined with K-nearest neighbor(KNN) imputation method (Troyanskaya *et al.*, 2001) over imperfect datasets with different missing rates.

Although a smaller percentage of missing data makes data imputation more precise, our algorithm is robust to increasing the percent of values missing. As the missing rate is increased, the conventional method loses accuracy dramatically; however, our algorithm maintains the accuracy in acceptable ranges over 90%. It is notable that our algorithm does not use any imputed values. The result means that the imputed value could mislead the clustering result seriously when the missing rate becomes high. This is because the usual imputation method which also use the correlation metric cannot use the information of the genes having missing entries in any sample of the dataset. For the imputation of a missing value of a target gene, the group of highly similar reference genes with no missing entries is required. The increase of the missing rate increase the number of less similar genes in this reference group and the resulting imputation value would mislead the clustering of the target gene. The result from iris data with missing entries supports that the proposed method

Table 3.3. The average silhouette coefficients of various clustering methods for the four microarray datasets.

Method	Datasets				Ave
	SRBCT	Colon	5 Tissue	Breast	
PM	0.22	0.20	0.62	0.20	0.31
K-means	0.15	0.23	0.48	0.18	0.26
SVD	0.17	0.20	0.49	0.16	0.25
Hierarchical	0.21	0.17	0.19	0.13	0.17
Diana	0.18	0.23	0.26	0.24	0.23
Fuzzy	0.20	0.12	0.05	0.13	0.12

Table 3.4. Performances of proposed method(PM) and K-means method combined with KNN imputation(KM) on the four microarray datasets with different missing rates.

Method		missing rate (%)				
		5	10	15	20	25
SRBCT	(PM)	99.2	97.8	97	96.33	95.4
	(KM)	98.8	97.18	96.5	96.1	95
Colon	(PM)	97.5	95.4	94	92.7	88.3
	(KM)	97.4	96.6	93.23	85	64.4
5 Tissue	(PM)	99.4	98	97.4	95.55	95
	(KM)	99.2	98	95.91	92	84
Breast	(PM)	99.5	98.4	97	96.7	95
	(KM)	98.9	97.6	97.1	94.91	94.2

which directly cluster the incomplete dataset is an appropriate alternative solution to overcome the deleterious effect of imputation.

We have also compared the methods in the microarray data. The overall validity score of the resulting cluster was higher in the proposed method (Table 3.3).

In these real world data, the performance of the methods fluctuated depending on the dataset and the type of methods; however, our algorithm showed a stably high performance. In the case of microarray data, it is not possible to be sure what the true clusters are. Therefore we checked the relative accuracy by comparing the clustering result of a method from an incomplete dataset with the clusters acquired by the same clustering method with a complete dataset. Table 3.4 shows that the performance of our clustering method over the microarray datasets with different missing rate. All compared methods used the KNN method for the imputation.

Our algorithm outperformed as the iris dataset for colon cancer (Colon) and five tissue (5 Tissue) datasets; in addition, it worked similarly well for SRBCT and breast cancer (Breast) datasets. Once again, the performance of K-means with imputation method was significantly changed depending on the datasets; however, K-means with our algorithm showed a consistently high performance.

4. Conclusion

This paper presents a new clustering method that does not require an imputation step to estimate missing values by combining the information of one dimensional (single sample) clusters. Surprisingly, this approach was comparable or superior with conventional methods in its performance for the complete dataset. In the test of the model dataset with missing entries, the advantage of this method was significant. The same clustering method (K-means) showed a huge difference in its performance depending on the use of imputation and our direct clustering method when the miss-

ing rate was high. It can be practically useful in saving the data of some accidental microarray experiments with high missing entries. In addition, our algorithm showed the superior performance for the real microarray data analysis. The performance of the methods fluctuated depending on the types of microarray datasets; however, our algorithm maintained a stable performance.

Our algorithm is a totally new alternative clustering approach for imperfect microarray data. We suggest that it is worthwhile to check the clustering result through this alternative way without any imputed data for the incomplete microarray data especially when the missing rate of the data is high. However, it is important to exercise caution when drawing critical biological conclusions from data that is partially calculated. The goal of this method is to provide an accurate way of clustering a microarray with missing values in order to minimally bias the performance of the microarray analysis methods. However, calculated data should be flagged where possible and its significance on the discovery of biological results should be assessed in order to avoid drawing unwarranted conclusions

References

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. and Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6745–6750.

Alter, O., Brown, P. O. and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling, *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 10101–10106.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

Gan, X., Liew, A. and Yan, H. (2006). Microarray missing data imputation based on a set theoretic framework and biological knowledge, *Nucleic Acids Research*, **34**, 1608–1619.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm, *Journal of the Royal Statistical Series C*, **28**, 100–108.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M. and Mark, R. (2001). Gene-expression profiles in hereditary breast cancer, *The New England Journal of Medicine*, **344**, 539–548.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley, New York.

Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. and Meltzer, P. (2001). Classification and diagnostic prediction of cancers using expression profiling and artificial neural networks, *Nature Medicine*, **7**, 673–679.

Kim, D. W., Lee, K. Y., Lee, K. H. and Lee, D. (2006). Towards clustering of incomplete microarray data without the use of imputation, *Bioinformatics*, **23**, 107–113.

Le, K., Mitsouras, K., Roy, M., Wang, Q., Xu, Q., Nelson, S. F. and Lee, C. (2004). Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data, *Nucleic Acids Research*, **32**, e180.

Ouyang, M., Welsh, W. J. and Georgopoulos, P. (2004). Gaussian mixture clustering and imputation of microarray data, *Bioinformatics*, **20**, 917–923.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of data clusters via the gap statistic, *Journal of the Royal Statistical Society: Series B*, **63**, 411–423.

Troyanskaya,O. G., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 520–525.