Approximation of Missing Values in DNA Microarray Gene Expression Data

* Amanjot Kaur
A.P. CSE
Global Institute
Amritsar

Sukhwinder Bir
Lect. CSE
Beant College of Engg.
and Tech., Gurdaspur

Reet Kamal
A.P. CSE
RIMT-MAEC
Mandi Gobindgarh

ABSTRACT
In the past few years, there has been a detonation of data in the field of biotechnology. Gene expression microarray experiments produce datasets with numerous missing expression values due to various reasons, e.g. insufficient resolution, image corruption, dust or scratches on the slides, or experimental error during the laboratory process.. To improve these missing values, many algorithms for gene expression analysis oblige a complete matrix of gene array values as input, such as K nearest neighbor impute method, Bayesian principal components analysis impute method, etc. Accurate estimation of missing values is an important requirement for efficient data analysis. Main problem of existing methods for microarray data is that there is no external information but the estimation is based exclusively on the expression data. We conjectured that utilizing a priori information on functional similarities available from public databases facilitates the missing value estimation. Robust missing value estimation methods are required since many algorithms for gene expression analysis entail a complete matrix of gene array values. Either genes with missing values can be removed, or the missing values can be replaced using prediction. Current methods for estimating the missing values include sample mean and K-nearest neighbors (KNN). Whether the accuracy of estimation methods depends on the actual gene expression has not been thoroughly investigated. Under this setting, we examine how the accuracy depends on the actual expression level and propose new method that provides improvements in accuracy relative to the current methods in certain ranges of gene expression.

Key Words: Clustering, DNA Microarray, Fuzzy Logic.

1. INTRODUCTION
Gene expression microarrays provide a popular technique to monitor the relative expression of thousands of genes under a variety of experimental conditions [1]. In spite of the enormous potential of this technique, there remain challenging problems associated with the acquisition and analysis of microarray data that can have a profound influence on the interpretation of the results.

Gene expression microarray experiments can generate data sets with multiple missing expression values. Unfortunately, many algorithms for gene expression analysis require a complete matrix of gene array values as input. Methods such as hierarchical clustering [1] and K-means clustering are not robust to missing data, and may lose effectiveness even with a few missing values. Methods for imputing missing data are needed, therefore, to minimize the effect of incomplete data sets on analyses, and to increase the range of data sets to which these algorithms can be applied [2]. There are several ways to deal with missing values such as deleting genes with missing values from further analysis, filling the missing entries with zeros, or imputing missing values of the average expression level for the gene (‘row average’).

Missing values can lead to erroneous conclusions about data and substitution of missing values may introduce inaccuracies and inconsistencies. These values can negatively

impact discovery results, and errors or data skews can proliferate across subsequent runs and cause a larger, cumulative error effect. As well, most analysis methods cannot be performed if there are missing values in the data. Missing values may prevent proper classification and clustering [3]

So the proper and more accurate prediction of Missing values remains an important step on the way to get better results. The goal of missing value is to represent a accurate data set of genes, species, or other taxa. A variety of approaches have been proposed for estimating missing values in DNA microarrays. Some of these methods are very complex and take a lot of time, while other having less accuracy. As a result, for any technique there is always a possibility to improve accuracy of estimating the missing values. In this paper we proposed a method using different t-norm, which given us better results on Missing Values.

The remainder of this paper can be described as follows: Next section contains a description of the methods used for Missing value estimation. In Section III the proposed methodology and section IV discusses the results of method applied on data sets. The paper ends with conclusions and future directions.

2. TECHNIQUES USED IN LITERATURE

2.1 Row Average or Filling with Zeros

Row average method is currently accepted method for filling missing data are filling the gaps with zeros or with the row average [7]. Row averaging assumes that the expression of a gene in one of the experiments is similar to its expression in a different experiment, which is often not true.

2.2 Singular Value Decomposition

In this method it is required to obtain a set of mutually orthogonal expression patterns that can be linearly combined to approximate the expression of all genes in the data set. The principal components of the gene expression matrix are referred as eigengenes [7].

$$A_{mxn} = U_{mxm} \Sigma_{mxn} V^{T}_{nxn}$$

Here matrix VT contains eigengenes, whose contribution to the expression in the eigenspace is quantified by corresponding eigenvalues on the diagonal of matrix. Identify the most significant eigengenes by sorting them based on their corresponding eigenvalues. The exact fraction of eigengenes for estimation may change. Once k most significant eigengenes from VT are selected then estimate a missing value j in gene i by Regressing this gene against the k eigengenes and use the coefficients of regression to reconstruct j from a linear combination of the k eigengenes..

2.3 Weighted K-Nearest Neighbors

Consider a gene A that has a missing value in experiment, KNN will find K other genes which have a value present in experiment, with expression most similar to A in experiments 2–N (N is the total number of experiments). A weighted average of values in experiment from the K closest genes is then used as an estimate for the missing value in gene A. Select genes with expression profiles similar to the gene of interest to impute missing values. The norm used to determine the distance is the Euclidean distance [2].

2.4 Linear Regression Using Bayesian Gene Selection

In Gibbs sampling method the Gibbs sampler allows us effectively to generate a sample X0,…,Xm ~ f(x) without requiring f(x). By simulating a large enough sample, the mean, variance, or any other characteristic of f(x) can be calculated to the desired degree of accuracy. In the two variable case, starting with a pair of random variables (X, Y), the Gibbs sampler generates a sample from f(x) by sampling instead from the conditional distributions f(x|y) and f(y|x). This is done by generating a "Gibbs sequence" of random variables. Bayesian gene selection: It uses a linear regression model to relate the gene expression levels of the target gene and other genes.

3. METHODOLOGY

The main idea of this work is to combine the accuracy and the effectiveness of the ensemble

clustering techniques based on random projections, with the expressive capacity of the fuzzy sets, to obtain clustering algorithms both reliable and able to express the uncertainty of the data.

Random projections have recently emerged as a powerful method for dimensionality reduction. Each ensemble is composed by 25 base clustering and each ensemble method has been repeated 20 times. In Table 1 data set is given. Regarding to ensemble methods based on random projections, projections are taken with bounded distortion, Theoretical results indicate that the method preserves distances quite nicely; however, empirical results are sparse. In random projection, the original d-dimensional data is projected to a k-dimensional (k<< d) subspace through the origin, using a random k × d matrix R whose columns have unit lengths. Using matrix notation where is the original set of N d-dimensional observations is the projection of the data onto a lower k-dimensional subspace.

$$X_{k*N}^{RP} = R_{k*d} X_{d*N}$$

The key idea of random mapping arises from the Johnson-Lindenstrauss lemma, if points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved. Random projection is computationally very simple: forming the random matrix R and projecting the d × N data matrix X into k dimensions is of order O(dkN), and if the data matrix X is sparse with about c nonzero entries per column, the complexity is of order O(ckN)[5].

Strictly speaking, (i) is not a projection because R is generally not orthogonal. A linear mapping such as (i) can cause significant distortions in the data set if R is not orthogonal. Orthogonalizing R is unfortunately computationally expensive.

Instead, in a high-dimensional space, there exists a much larger number of almost orthogonal than orthogonal directions. Thus vectors having random directions might be sufficiently close to orthogonal, and equivalently would approximate an identity matrix.

The choice of the random matrix R is one of the key points of interest. The elements of R are often Gaussian distributed, but this need not be the case. Gaussian distribution can be re- placed [5] by a much simpler distribution. Practically all zero mean, unit variance distributions of would give a mapping that. In this thesis work random projection method is used to get different views of an original matrix.

$$r_{ij} = \sqrt{3} \begin{cases} +1 & \text{With probability } \frac{1}{6} \\ 0 & \text{With probability } \frac{2}{3} \\ -1 & \text{With probability } \frac{1}{6} \end{cases}$$

Data Set

TABLE I
DLBCL – FL DATA SET

310.54	144.26	0	118.98	273.85
120.49	57.02	61.38	96.9	122.6
112.24	147.09	219.47	234.68	176.05
1391.89	3241.79	4380.8	4376.86	0
115.83	85.87	162.04	72.81	79.15
524.27	395.935	610.465	0	574.74
338.87	143.38	341.12	167.29	84.68
54.87	94.55	118.19	106.33	65.32
22.91	79.785	32.065	47.41	33.78
165.67	887.68	322.14	423.92	313.11
49.49	0	130.72	39.68	311.86
198.66333	190.75	105.2766	61.09667	150.02667
2581.18	557.45	683.41	607.07	0
328.12	222.72	326.64	382.3	284.56
71.72	36.65	20	20	36.85
52.36	146.79	97.19	103.25	76.11

Design of Algorithm

The algorithm for estimating the missing values with DNA microarray gene expression was designed after studying various algorithms that can used for estimation techniques. This algorithm has following steps:

Input

- A data set X = { $x_1, x_2, \ldots x_n$ }, stored in a d × n D matrix.
- An integer k (number of clusters)
- An integer c (number of clustering)
- The fuzzy k-means clustering algorithm C

- Procedure the realizes the randomized map μ
- An integer d' (dimension of the projected subspace)
- A function τ that defines the t-norm

Begin

- For each i, j ∈ {1, . . . , n} do $M_{ij} = 0$
- Repeat for t = 1 to c
- R_t = Generate projection matrix (d', μ)
- $D_t = R_t .D$
- [IDX,C]=kmeans(D_t ,n)
- For each i, j ∈ {1, . . . , n}
- $M_{ij}^{(t)} = \sum_{s=1}^{k} \tau(C_{si}^{(t)}, C_{sj}^{(t)})$
- End Repeat
- $M^c = \frac{\sum_{t=1}^{c} M^{(t)}}{c}$
- $< A_1, A_2, \ldots, A_k >= \text{kmeans}(D_t, n$

End

The final Clustering C=<A_1, A_2…A_k> and cumulative similarity matrix M^c .Inside the mean loop the procedure Generate projection matrix produces a d' × d R_t matrix according to a given random map μ, that it is used to randomly project the original data matrix D into a d' × n D_t projected data matrix. Next, the fuzzy k-means algorithm with a given fuzziness is applied to D_t and a k-clustering represented by its C membership matrix is achieved. Hence the corresponding similarity matrix $M_{ij}^{(t)}$is computed, using a given t-norm. Next the" cumulative" similarity matrix M^c is obtained by averaging across the similarity matrices computed in the main loop. Finally, the consensus clustering is obtained by applying the fuzzy k-means algorithm to the rows of the similarity matrix M^c. The Consensus clustering step is performed by applying the fuzzy-k-means clustering to the rows of M^c. Indeed i^{th} row of M^c represents the "common membership" to the same cluster of the i^{th} example with respect to all the other examples, averaged across multiple clustering. In this sense the rows can be interpreted as a new "feature space" for the analyzed data.

In this work one DNA microarray data sets is used i.e. (DLBCL-FL data set) is composed by tumor specimens from 58 Diffuse Large BCell Lymphoma (DLBCL) and 19 Follicular Lymphoma (FL) patients. For each ensemble projections are randomly repeated 20 times, and each time fuzzy ensembles is composed by 20 base clustering. Table 2 shows the results after applying the algorithm on the original data. Using these results fuzzy k-mean method can be compared with other methods for accuracy. To Test the performance of proposed method, these results are compared with other existing results of the other methods of clustering algorithms

TABLE 2
Resultant Data SET

43.4294	14.1795	8.8683	8.3586	5.5838
78.6295	170.7405	212.5668	214.1777	13.3964
105.6501	175.0164	175.9064	177.3915	163.6902
98.2399	169.9005	158.0186	161.6823	150.5259
74.384	158.3449	202.9421	200.5964	243.5543
96.1167	176.3022	215.3371	217.3607	215.0804
100.8813	101.2433	86.8734	86.883	75.492
91.0148	179.8443	112.091	11.095	10.6686
96.2893	176.7901	211.4245	207.567	204.4082
81.6819	169.5082	215.128	213.9146	249.9721
86.518	163.8384	202.3706	207.0531	202.0059
95.7827	186.7053	190.258	186.067	180.178

4. RESULTS AND DISCUSSION

Table.3 shows the compared numerical results of the experiments on the DLBCL-FL data set. Fuzzy algorithm method obtained better results with respect to other methods. This table represents the compared results of proposed fuzzy min ensemble algorithm with other algorithms. According to [4] if median error has higher value, that algorithm has better accuracy. In this table, there is higher value of median error in fuzzy algorithm. That shows that this algorithm has better accuracy as compared to other algorithms.

Five previous methods are used to compare the performance of proposed method in this thesis work. One advantage of the proposed method is that it makes most use of the information from the

original data set. Random projection scheme raises the estimation performance notably, which contributes to the best performance of proposed method among other methods. KNN method linearly combines the similar genes by weighting the average values of them. In KNN method a function T-norm is used. T-norm is kind of binary operation used in the framework of probabilistic metric spaces and in multi-valued logic, specifically in fuzzy logic. They are a natural interpretation of the conjunction in the semantics of mathematical fuzzy logics and they are used to combine criteria in multi-criteria decision making.

TABLE 3
COMPARISON CHART OF DLBCL-FL GENE EXPRESSION DATA

Algorithm	Median Error
Proposed Fuzzy Min Ensemble	0.2572
Fuzzy Max	0.2208
Fuzzy Alpha	0.2208
Max Max	0.2468
Max Alpha	0.2468
Rand Clust	0.1039

A triangular norm (abbreviation t-norm) is a binary operation on the interval [0,1] satisfying the following conditions

- $T(x,y) = T(y,x)$ (commutative)
- $T(x,T(y,z)) = T(T(x,y),z)$ (associativity)
- $y \leq z \Longrightarrow T(x,y) \leq T(x,z)$ (monotonicity)

$T(x,1) = x$ (Neutral element 1).

5. CONCLUSION AND FUTURE SCOPE

The experiment results show that fuzzy algorithm provides better results than existing algorithms. Present work concludes that random projection is a good alternative to traditional, statically optimal method for dimensionality reduction. This thesis work presented new and promising experimental results on random projection in dimensionality reduction of high-dimensional real-world data sets. When comparing different methods for dimensionality reduction, the criteria are the amount of distortion caused by the method and its computational complexity. The algorithm used 58 tumor species samples from DLBCL-FL data set. Algorithm also used a good approach that is Random Projections, in which high dimension matrix is projected to low dimension matrix to obtain better results. These results indicate that random projection preserves the similarities of the data vectors well even when the data is projected to moderate numbers of dimensions; the projection is yet fast to compute. Experiment with multi label genes to show more clearly the effectiveness of the proposed approach to analyze the structure of unlabeled data when the boundaries of the clusters are uncertain.

6. REFERENCES

[1] F.Luo, K.Tang, L.Khan . "Hierarchical clustering of gene expression data" in Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium, 2003, pp. 328-335.

[2] 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. 17,2001, pp.520- 525.

[3] B.M.Nogueira, T.R.A.Santos, L.E.Zarate,"Comparison of Classifiers Efficiency on Missing Values Recovering: Application in a Marketing Database with Massive Missing Data" at Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium, 2007, pp. 66-72.

[4] A.Bertoni, G.Valentini "Ensembles based on random projections to improve the accuracy of clustering algorithms" in Neural Nets, WIRN

2005. Volume 3931 of Lecture Notes in Computer Science., Springer,2006, pp. 31–37.

[5] E. Bingham and M. Heikki "Random projection in dimensionality reduction: Applications to image and text data", International Conference on Knowledge Discovery and Data Mining, San Francisco, California,2001,Vol. 18, pp 245-250.

[6]A. Bertoni and V. Giorgio "Fuzzy ensemble clustering for DNA microarray data analysis", Lecture Notes in Computer Science,,2007 Vol. 3931, pp 537-543.

[7] 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,1. 1. 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, 2000, pp. 503- 511