VOL. 12, NO. 2, JANUARY 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

AN INTEGRATED SEMI-SUPERVISED CLUSTERING MODEL FOR TIME COURSE GENE EXPRESSION DATA

Peter Juma Ochieng[1] and Taufik Djatna[2]
[1]Department of Computer Science, Institut Pertanian Bogor University, Bogor, Indonesia
[2]Department of Agro Industry, Institut Pertanian Bogor University, Bogor, Indonesia
E-Mail: peter26juma@gmail.com

ABSTRACT

Clustering the time course data using basic conventional clustering methods often, present computational challenges and most algorithms are porn error when dealing with such data structures. Thus, the aim of this study is to introduce an integrated semi-supervised model for clustering time course gene expression data. The proposed model implement four series approximation to account for the periodic gene expression; AR(1) mixed random effect to account for the auto correlated data structure for time course gene expression and rejection controlled EM algorithm to minimize the computational cost during m-step. The interest of the proposed method is illustrated by its application to yeast cell life cycle dataset. Simulation results indicate the proposed method to cluster the various genes expression to their correct profiles. Further empirical comparison indicates the proposed method to outperform the HMRF-Kmean with 0.154 error rate; 0. 785 rand index and 0.592 adjusted rand index. Therefore, integrating the Fourier series approximation, AR (1) random effect model and rejection controlled EM algorithm the proposed model provides a more reliable and robust method for clustering time-course data since the model allows for the correlation among observations at different time points.

Keywords: gene expression, clustering model, AR (1), fourier series.

INTRODUCTION

Microarray experiments are widely used to measure gene expression profile of a given time course or different cellular conditions. Therefore, to understand biological meaning of such expression profile over a given time it is critical to initially group the expressed genes bases on the similarity of their expression profiles [1][2].

To a group such genes with similar expression profile, several clustering methods have been proposed including hierarchical clustering, self-organizing maps, and k-means clustering have been used to decipher the unique structures of gene expression patterns [3].

In addition, artificial intelligence methods like neural networks and Bayesian networks have recently been proposed to cluster time course gene expression data however, the size and temporal nature of the time course gene expression data tend to make the two methods less efficient by increasing the convergence time due to numerous hidden layers and training process [4].

To overcome such obstacle, other methods such as multi-variant Gaussian models have been proposed to model the time course data in order to account for the correlated data structure. However, considering the temporal time course data those methods often deliberately ignore the time order for gene expression [5].

To account for the temporal nature of the time course data methods such as Cluster Analysis of Gene Expression Dynamic (CAGED) have been proposed to cluster time course gene expression data. The method utilizes the auto regression approach to model the time series gene expression data, however, the great challenge using this method is that the Markov property is unlikely to hold since this model often requires stationary [6].

Semi-supervised clustering method such as Hidden Markov Random Field (HMRF) has been proposed for clustering gene expression data [7]. The algorithm utilizes an EM algorithm to compute and update cluster membership for gene expression profiles, however, this method often uses stationary making it unsuitable for clustering the time course gene expression data. Furthermore, the method requires appropriate random mixture effect model initially transform of the time course gene expression data. Another drawback when clustering time course data using HMRF K means is that the method often requires high computational time during the EM-step this occasionally make the algorithm unstable and prone to error especially when handling huge time course data [8].

to overcome those aforementioned obstacles we proposed an integrated semi-supervised clustering model that utilize the Fourier expansion, first order autoregressive mixed random effect model AR(1) and rejection controlled EM algorithm (RCEM). Therefore, the aims of this study were to implement Fourier expansion to account for periodic gene expression, AR (1) mixed random effect model to account for the correlated structure of time course gene expression data and to introduce a RCEM algorithm to minimize the computational cost during the M-step.

PRELIMINARIES

Several methods have been advocated by researchers to cluster gene expression data with the focus on the discovery the unique gene expression pattern at different cellular condition and time course [7]. Techniques which incorporate mixture models have been on forefront for clustering time course gene expression data to discover unique gene expression profile for given biological phenomenon [8].

Method like a bi-clustering approach has been used to identify a co-expressed set of genes in microarray gene expression data as part of the experimental conditions [9]. Nevertheless, this method often allows for overlapping gene clusters and reveals subtle gene clusters; furthermore, this method does not consider the independence of the multi-condition of gene expression and time series of gene expression data [10]. Therefore, using this method time points are clustered independently and the relationship between time points are ignored rendering the method inappropriate for clustering time course gene expression [11].

Recently gene set analysis (TCGSA) has been proposed to cluster the predefined groups of genes in the analysis of gene expression data in cross-sectional studies [12]. This method is an extension of gene set analysis to longitudinal data which often relies on the random effect to model and maximum likelihood estimates for clustering the time course gene expression data, however, the method often relies on the use of available repeated measurements when handling unbalanced data or missing data at random (MAR) measurements [13]. Despite TCGSA based on the assumption driven method to identifies a priori defined gene sets with significant expression variations over time the method don't implement the Fourier series approximation to account for periodic time expression within gene sets[14].

Several computational methods have been developed for clustering gene from microarray data. The major focus has been to develop methodologies to derive the formulation to account for the temporal patterns of the gene expression, especially in time course gene expression data. Application of methods such as principal component analysis (PCA) and mixture model have been widely utilized to identify the differential expression of genes over time [15]. For instance, finite mixture models have been proposed to model the distribution of random phenomena such as continues normal multi-variant data from nature [16].

The multi-variant normal mixture model has been used to detect different gene expression profiles based on the assumption that there is no replication of any specifically identified entity and all the entity observed are independent of one another, this makes the mixture models inadequate to handle time course data [17]. However, both assumptions will not hold for clustering gene profile since the gene profiles are measured over time and all genes are independently distributed. Since the correlated data structure can be included in the normal mixture model a reliable component covariance data matrices, the challenge is often how to model such specific data structure [18].

Method like EM-based mixture analysis with random effects has been developed to cluster correlated data that may be replicated. This method utilizes the linear mixture model to account for the correlation structure between the variables to enable the correlations among the observations. Based on the conditional issue specific the random effect can be formulated especially during the EM-steps in closed form [19]. However, the basic EM algorithm often experiences computational drawbacks in the E-step by carrying out unnecessary time-consuming Monte Carlo methods [20]. Therefore, the probabilistic clustering of the genes into g components can be achieved by estimating the posterior probabilities components of membership given the profile vectors and estimating specific random effect [21][22].

Fourier series approximations are often used approach to model periodic gene expression, thus its application may facilitate detection of periodic gene expression for various organisms including yeast and human cells. Especially when the studied genes are periodically regulated therefore Fourier series approximation can be used to approximate their time dependence [23]. Furthermore, in this study, we proposed to design a rejection-controlled EM algorithm for estimation. Where the E-step is followed by a rejection-controlled sampling step to eliminate functional observations, whose posterior probabilities of belonging to a particular cluster is negligible for calculation in the subsequent m-step thereby reducing the computational cost and error.

METHODS

To achieve the objectives of this study, we initially formulate an autoregressive random mixed effect model using Fourier expression for periodic expression and first order auto regression model AR (1) to account for the overlapping time coursed data structure.

Secondly, we implement the rejection controlled EM algorithm to set the threshold for gene-to-cluster membership probabilities. Note that in this study we assume (a) at any particular entity identified there is no replication and (b) all the entities are independent of one another.

Since genes are often regulated, we applied Fourier series approximation to model the periodic gene expression. Therefore, we express a general k^{th} order of Fourier series expansion as:

$$gk(t) = \sum_{j=1}^{k} [a_j \cos(2\pi jt/\omega) + b_j \sin(2\pi jt/\omega)] \quad (1)$$

Modeling time course gene expression

The initial step of this study is to model the time course gene data using AR (1) for efficient clustering of the time course data by basic EM algorithm.

Therefore, to model gene expression data, let X denotes the model matrix and β the associated vector for regression coefficients for the mixed effect. Linear mixed model with vector y_j for the j^{th}gene and the h^{th} for conditional membership of the components mixture is adopted [19] and presented as:

$$y_j = X\beta h + Z_1 u_{jh} + Z_2 u_h + \varepsilon_{jh} (j = 1, \ldots n) \quad (2)$$

where β_h are unknown parameters given as 2k+1 whereas the random effects are a_0, a_1, a_k, b_1,….b_k and

$u_{jh}=u_{jh1},\ldots u_{jhm}$ and $v_h=(v_{h1},\ldots v_{hm})^T$ where m is the number of time points in the equation (1) and (2). Z_1 and $Z_2$are $m\times m$ identity matrices. In our formulation we assume ε_{jh} and v_h to be normal distributed and independent $N(0, D)$ and, $N(0, \Omega)$ independent of u_{jh}. Therefore, to account forthe time dependent random gene effects first-order, we adopt autoregressive correlation structure for the gene profiles, thereby presenting the normal distribution $N(0, \theta^2 A(\rho)$ by:

$$A(\rho)=\frac{1}{1-\rho^2}\begin{pmatrix}1 & \rho & \ldots & \sigma^{m-1}\\ \rho & 1 & \ldots & \sigma^{m-2}\\ \vdots & \vdots & \ldots & \vdots\ \vdots\\ \sigma^{m-1} & \rho^{m-2} & \ldots & 1\end{pmatrix} \quad (3)$$

$$A(\rho)^{-1}=(1+\rho^2)I-\rho J-\rho^2 K \quad (4)$$

therefore, inverse of $A(\rho)$ is :

$$\frac{A(\rho)^{-1}}{\partial\rho}(A\rho)=-\frac{2\rho}{1-\rho^2} \quad (5)$$

where *J, K, and I* are *m x m*matrices and Iis identity matrix and *J* is the sub-diagonal entries with ones and zeros on other entries, K the principal diagonal takes value 1 and zeros elsewhere at the first and last element.

To compute maximum likelihood estimation of parameters, we will use equation (3) and (4). Based on the assumption (a) and (b) we assume autocorrelation covariance structure at each time point and a larger autocorrelation residual structure. In this step, we adopted g-component mixture with the probability density function (pdf) as:

$$f(y|\Psi)=\sum_{h}^{g} p_h f_h(yj|\beta_h\Omega_h,\theta_h^2,A_h,D_h) \quad (6)$$

where f_h is the component of probability density function of the multivariate normal distribution with mean vector $X_h\beta_h$ while the covariance matrix expressed as:

$$\theta_h^2 Z_1 A_h Z_1^T + Z_2 D_h Z_1^T + \Omega_h$$

For unknown vector parameters Ψ we estimated the maximum likelihood using the proposed rejection controlled EM algorithm.

Likelihood estimation via RCEM algorithm

In this step the maximum likelihood is estimated using RCEM here the observed data vector $(y=y_1,y_2,..,y_n)^T$ is considered as augmented components with unobserved labels, $Z_1,Z_2,..,Z_n$ of $y_1,y_2,..,y_n$. Thus, Z_j is the g-dimensional vector with h^{th} element Z_{jh} equal to 1 if $y_1$is the result of h^{th} component of the mixture otherwise zero. In this case, the hidden values are considered as missing values or data thereby they are included and considered as incomplete data vector. Thus the random effect vectors u_{jh} and v_h(j=1,2,3,…n: h=1,2,3,..,g) are considered as missing values and included as a complete data vector. Thus, log-likelihood $l_c$of the complete data is representedwith the summation of the four terms given as $l_c=l_1+l_{2+}l_3+l_4$ expressed as:

$$l_1=\sum_{h=1}^{g}\sum_{j=1}^{n} Z_{jh}\log(ph) \quad (7)$$

where l_1 is the logarithm probability of the component labels Z_{jh}, and whereas l_2 is the logarithm density function of *y* conditional on u_{jh}, v_h and Z_{jh}=1, and l_3 and $l_4$is the logarithm of the density function of *u* and *v*, respectively, therefore if $Z_{jh}=1$ then:

$$l_2=\sum_{h=1}^{g}\sum_{j=1}^{n} Z_{jh}\,(\mathrm{mlog}(2\pi)\log|\Omega_h|+\varepsilon_{jh}^T\Omega_h^{-1}\varepsilon_{jh} \quad (8)$$

$$l_3=\sum_{h=1}^{g}\sum_{j=1}^{n} Z_{jh}\,(\mathrm{mlog}(2\pi\theta_h^2)\log|A_h|+\theta_h^{-2}A_h^{-1}u_{jh} \quad (9)$$

$$l_4=\sum_{h=1}^{g}\sum_{j=1}^{n} Z_{jh}\,(\mathrm{mlog}(2\pi)\log|D_h|+u_h^T D_h^{-1}v_h \quad (10)$$

Maximum likelihood penalization

For maximization of the log likelihood l_c in (7) we perform decomposition for each values of$l_1, l_2, l_3,$ and l_4 using a maximum penalization likelihood approach. Thus, using Henderson's likelihood penalization we express the complete data as:

$$l_c=constant-\left(\sum_{h=1}^{g}\sum_{j=1}^{n} Z_{jh}\log(ph)+l_2+l_3+l_4\right) \quad (11)$$

The next step is to compute EM of equation (12). Here we calculate the probability of each gene belonging to a given cluster base on the expression:

$$w_{jh}=\frac{p_h\varphi(y_j\mu_h(X_i),\Sigma_h}{\sum_{l=1}^{h}p_l\varphi(y_j\mu_l(X_i),\Sigma_l} \quad (12)$$

where$\Sigma_h=A_h\Omega_h D_h+\sigma^2 I$, and φ represent the Gaussian density function. In the M-step, conditional minimization is expressed by equation:

$$-\sum_{h=1}^{g}\sum_{j=1}^{n} w_{jh}\log(ph)+\frac{1}{2\sigma^2}\left(\sum_{h=1}^{g}\sum_{j=1}^{n} Z_{jh}\log(ph)+l_2+l_3+l_4\right)$$

However, since the M-step involves minimization thus equation (13) is reduced to:

$$p_h=\frac{1}{n}\sum_{j=1}^{n} w_{jh}\,for\,h=1,\ldots.H \quad (13)$$

For variance estimation, we adopted the equation below to measure the error:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{h=1}^{g}\sum_{j=1}^{n} w_{jh}(y_j - \mu_h(X_j) - A_h\varepsilon_{jh})^T(y_j - \mu_h(X_j) - A_h\varepsilon_{jh})$$

Thus the probability that particular gene belongs to each cluster given all the parameters in the model are express by equation:

$$P(gene_j \in h) = \frac{p_h N(\mu_h, \int_h)}{p_1 N(\mu_1, \int_1) + \cdots + p_h N(\mu_h, \int_h)}$$

Using the equation above we asymptotically minimizes the discrepancy between the true and estimated expression profiles by estimating the error of variance($\hat{\sigma}^2$) of the data based on RMSE. Thus cluster parameterP_h is updated until convergence however iteration process is often unstable using basic EM:

$$P_h = \frac{[\sum_{j=1}^{n} P(gene_j \in h) + \alpha_h]}{(n + \sum_{j=1}^{n}\alpha_h)}$$

Rejection-controlled algorithm (RC)

To stabiles, the iteration process we adopted RC algorithm to set gene-to-cluster membership probability threshold value c. Thus, membership probability for rejection control step is expressed as:

$$w_{jh}^* = \begin{cases} w_{jh} & if \quad w_{jh} > c \\ c \quad with probability & w_{jh}/c \quad if w_{jh} \le c \\ 0 \quad with probability & 1 - w_{jh}/c \quad if w_{jh} \le c \end{cases}$$

where w_{jh}^* is normalized w_{jh} given by $w_{jh}^* = w_{jh}/\Sigma_h w_{jh}$ to check the functionality of the proposed RC we set a low threshold value (c = 0.05) for gene-to-cluster membership probabilities. Thus, genes in cluster membership greater than threshold value c are unaffected, whereas genes with gene-to-cluster membership less than c as assigned probability zero or probability c for a particular cluster at hand. This is expressed as $1 - P(gene_i \in h)/c$ and $P(gene_i \in h)/c$ respectively thus setting a 'low-probability almost zero the summation cost in (11) is greatly reduced.

Determination of number of clusters

The success of the proposed mixed-effect model depends on the selection of number of cluster. Therefore, to determine the number of clusters we adopt the Bayesian Information Criterion (BIC) to estimate approximates Bayes factor:

$$BIC(C) = 2L(C) - m_C \log n$$

where L(C) is the maximization log likelihood for the model with C cluster, m_C is the number of independent parameters to be estimated in C cluster model and n is the sample size. Although, the regularity condition for BIC do not hold for mixture models we opted to apply it due to its wide application in most mixture models for analysis of microarray data.

RESULTS AND DISCUSSIONS

To check the performance of the proposed model we present a simulation analysis using datasets of 384 genes of Yeast cell cycle. Initially, we assigned "labels" which we considered as the "Main Group" thus we generated five "main groups" namely G1, late G1, S, G2, and M. Based on that we compared the cluster results of the all the phases (labels).

According to figure 1, by setting the number of cluster four (n=4) all the gene have similar expression profiles in each cluster except in cluster 2 where there is greater variation in expression by some genes. Furthermore, cluster 4 consisted of majority (153 genes) of genes with the maximum expression peak at 100 time point. Cluster 1 consisted of 28 genes with maximum expression peak at 25 time point. Cluster 3 consisted of 37 genes with the maximum peak expression at 25 and 100 time points. Cluster 2 consisted of 19 genes unique expression peak since similar genes were expressed at same time points this shows a significant correlation.

When we set, n = 5 we obtain five clusters as illustrated in figure 2. According to the figure cluster 2, 3 and 4 genes have similar expression pattern whereas genes in cluster 1 and 5 are similarly expressed. Those patterns clearly indicate the proposed method to efficiently model the time course data to significantly discover the correlation between genes expression and time.

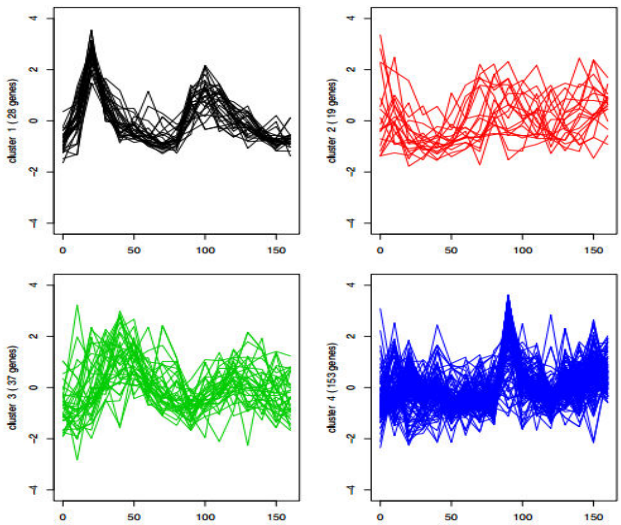

Figure-1. Gene expression profiles of the four cluster for the yeast data.

Implementation of the proposed RC algorithm in the EM step, the computational cost DURING M-step is greatly reduced by adjusting the threshold value for gene-to-cluster membership hence minimizing the computation of log likelihood in the m-step. From the simulation setting the threshold value c= 0.05 for gene-to-cluster membership the performance of the RCEM algorithm is optimized. However, adjusting the threshold value to c =1.0 the algorithm is reduced to Monte Carlo EM algorithm reducing the clustering performance.

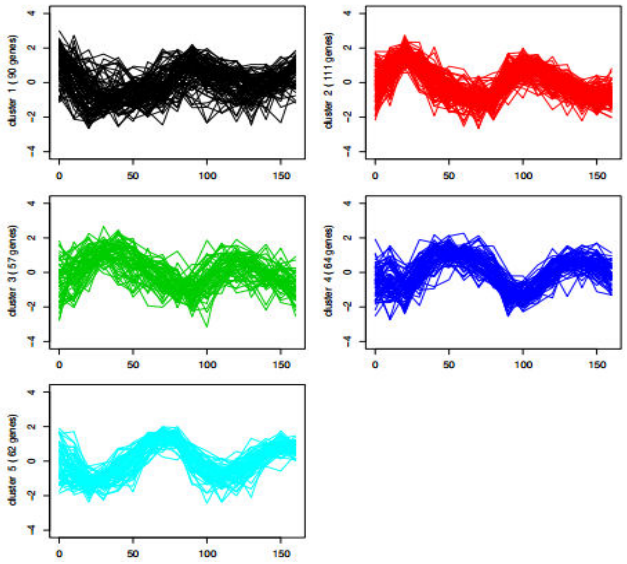

Figure-2. Gene expression profiles of five clusters for the yeast data.

Further adjustment of the threshold value c= 0.0 reduce the RCEM algorithm to original EM algorithm. Thus from this simulation, the optimal threshold value was reached at c =0.05 making the algorithm to be fast and robust since at this threshold value assigning the gene-to-cluster membership probabilities optimized.

This implies that the genes falling within this cluster membership greater than the set threshold value were unaffected. Thus, setting the threshold values close to zero the summation cost in the M-step is greatly minimized. In this analysis, we initial set the threshold value 1 after which we it was adjusted gradually towards zero to attain the original EM algorithm. Therefore, to achieve this we run the algorithm with multiple chains to avoid local optimal. Since the likelihood iteration for original EM algorithm keeps on increases gradually and only stop when the likelihood iteration is static , the proposed RCEM algorithm adopt Gibbs sampler on the intuition that when the likelihood is at stationary for several consecutive iterations, we stop and select the estimated highest likelihood.

The integrated semis-supervised model incorporate a mixture model with AR (1) random effect and rejection controlled (RC) algorithm to cluster time course gene expression data. By implementing Fourier expansion, we model the periodic patterns and to account for autocorrelation observations at different time points we use the autocorrelation variance structure. AR (1) account for the correlation between genes and between time points for precise clustering of the time course data. Simulation analysis indicates an improved accuracy of clustering if the autocorrelation of the time-dependent genes is accounted for over a given time points since we assume that periodic time is fixed.

Setting cluster effects zero (v=0) and random effects u of the each autocorrelated genes our model become flexible for clustering the time course gene expression data. However, when both random effects u and v are set zero, our model is reduced to normal mixture regression model. For likelihood estimation, we adopted penalized Henderson's likelihood. The rejection controlled EM algorithm implemented to reduce the expensive computational cost for large scale data. Moreover, the Bayesian information criterion used to determine the number of the cluster.

For comparison, we measure the misclassification error rate, adjusted rand index and rand index when similar yeast dataset clustered using proposed AR (1)-RCEM algorithm with HMRF-Kmeans. We considered rand index, which is the percentage of concordance pairs over all possible data pairs. An adjusted rand index, which is the maximum value when two clustering results are the same and the expected value, is equal to zero when two clustering results are independent. In this case, the adjusted rand index and rand index assess the level of agreement between the true cluster and the partition; therefore, the larger the adjusted rand index the higher the level of cluster agreement.

To compare AR (1)-RCEM and HMRF-Kmeans we considered adjusting parameters including p, c, a_0, a_1, b_1, θ^2, ρ and σ^2. According to Table-1, it can be seen the proposed model outperform HMRF kmeans with 0.010 error rate, 0.014 rand index and 0.28 adjusted rand. However, for HMRF Kmean experience there was an increase in error rate and adjusted rand index especially where the gene expression profiles are correlated and consist of a big data matrix.

Table-1 illustrate the comparison analysis of bias and RMSE of AR (1)-RCEM and HMRF-Kmeans based on optimal parameter adjustments of p, c, a_0, a_1, b_1, θ^2, ρ and σ^2 values. In this study, we consider misclassification parameter including error rate, rand index and adjusted rand index estimations to estimate the bias and RMSE for both algorithms.

Table-1. Bias and RMSE on simulated dataset of 384 genes yeast cell life cycle.

	ADJUSTED PARAMETERS		MEAN	RMSE	SD
AR(1)-RCEM	p (0.58,0.10,0.35) a_0(0.30,0.10,0.02) b_1 (0.50,0.10,0.09) σ (0.60,0.90,0.01) ρ (0.40,0.60,0.80) θ^2 (1.30,0.13,0.65) c (0.00,0.05,1.00)	Error Rate	0.154	0.153	0.010
		Rand	0.785	0.201	0.014
		Adjusted Rand	0.592	0.410	0.028
HMRF-Kmaen	p (0.58,0.10,0.35) a_0(0.30,0.10,0.02) b_1 (0.50,0.10,0.09) σ (0.60,0.90,0.01) ρ (0.40,0.60,0.80) θ^2 (1.30,0.13,0.65)	Error Rate	0.163	0.168	0.013
		Rand	0.794	0.219	0.017
		Adjusted Rand	0.561	0.437	0.028

Figure-3 illustrate the comparison between AR (1)-RCEM and HMRF Kmean. To check the reliability of the proposed method the threshold value for the gene-to-cluster membership was an adjustment at a range of 0.0 to 1.0. According to the analysis, it can be seen when the

threshold value is set 0.05 the error rate, adjusted rand and the rand index greatly reduced for AR (1)-RCEM algorithm hence the gene-to-cluster membership in the maximization step (M-step) is optimized.

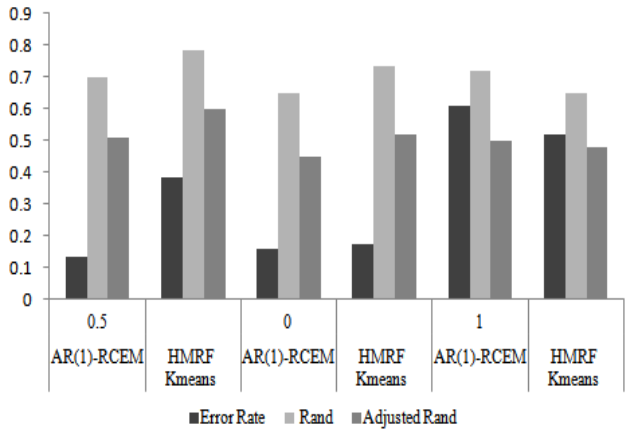

Figure-3. Comparison analysis of AR (1)-RCEM and HMRF-K means

Furthermore, when the threshold value is adjusted to 0.0 the proposed AR (1)-RCEM is reduced to original EM algorithm thus the algorithm is prone to error and high computational cost during the M-step this results to similar performance with HMRF-Kmean, which often utilize the classic EM algorithm during the exceptional and maximization step. Moreover, when the threshold value is set 1.00 the error rate, adjusted rand and rand index is increased compared to HMRF-Kmean this because by setting threshold 1.00 the RCEM algorithm is reduced to Monte Carlo EM algorithm.

CONCLUSIONS

Our proposed integrated semi-supervised clustering model for time course gene expression data successfully implements Fourier expansion to account for periodic time and AR (1) mixed random effect model account for autocorrelation between observed genes at different time points. The introduction of rejection control EM algorithm, we overcome the aforementioned obstacles in the M-step for original EM algorithm hence we minimize the computational cost and associated error. Simulation results demonstrated the proposed AR (1)-RCEM to be accurate and robust to cluster time course gene expression data further comparison analysis indicate AR (1)-RCEM to outperform HMRF-Kmeans with 0.154 error rate; 0. 785 Rand Index and 0.592 adjusted Rand Index.

ACKNOWLEDGEMENTS

The authors express their gratitude to the IPB University for its kind support towards this research and the computer science department, faculty of mathematics and natural science for insightful comments and suggestions for making this paper and guide on how to implement the preparation and search related to this study.

REFERENCES

[1] N. Coffey, J. Hinde, "Analyzing time-course microarray data using functional data analysis a review". Statical Application Genetic Molcule. 10, 2011.

[2] W. E. Johnson, C. Li, A. Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". Biostatistics. 2007; 8(1):118-127.

[3] P. Tamayo , D. Slonim , J. Mesirov." Interpreting Patterns Of Gene Expression With Self-Organizing Maps: Methods And Application To Hematopoietic Differentiation" Practical National Academy Science U S A. 1999; 96:2907 -12.

[4] C. Hafemeister, I. G. Costa, A. Schonhuth, A. Schliep, "Classifying short gene expression time-courses with bayesian estimation of". Bioinformatics. 27:946-952, 2011.

[5] B. Shahbaba, R. Tibshirani, CM Shachaf, SK. Plevritis. "Bayesian gene set analysis for identifying significant biological pathways." Journal of the Royal Statistical Society Series Applied statistics. 2011. 60 (4):541–557.

[6] D. M.C. Souto, I.G Costa, D. Araujo. "Clustering cancer gene expression data: A Comparative Study. BMC Bioinformatics. 2008; 9:497. [Pubmed: 19038021]

[7] S. Smith, Y. Zhang and M. Brady "Hidden markov random field model and segmentation of brain mr images". IEEE transactions on medical imaging, Vol. 45-57, no. 20(1), 2001.

[8] H. Jacqmin-Gadda, S. Sibillot, C. Proust, J. M. Molina, R. Thiébaut, D. C. Koestler, C. J. Marsit, B. C. Christensen, "Semi-supervised recursively partitioned mixture models for identifying cancer subtypes.". Bioinformatics. 26:2578-85, 2010.

[9] Y. Cheng and G. Church, "Biclustering of expression data." In Proceedings of 8th International Conference onIntelligent System for Molecular Biology (ismb), Vol. 93-103, 2000.

[10] M. J. Nueda, J. Carbonell, I. Medina, J. Dopazo, A. Conesa. "Serial expression analysis: A Web tool for The Analysis of serial gene expression Data". Nucleic Acids Research. 2010. 38: W239– W245.

[11] L. F. T. Scharl, B. Grun. "Mixtures of regression models for time-course gene expression data: evaluation of initialization and random effects," Bioinformatics, vol. 370-377., no. 26, 2010.

[12] H. Maciejewski, "Gene set analysis methods: statistical models and methodological differences". Briefings in Bioinformatics. 2014. 15(4):504–518.

[13] A.P Dempster,N.M Laird, D.B Rubin. "Maximum Likelihood From Incomplete Data Via The EM algorithm". British Royal Journal of Statistic Society . 1977.

[14] M. Berk, C. Hemingway, M. Levin, G. Montana, "Longitudinal analysis of gene expression profiles using functional mixed-effects models". Advanced Statistical Methods for the Analysis of Large Data-Sets. Springer. 2012. p. 57-67.

[15] T. Scharl, B. Grü, F. Leisch. Mixtures of regression models for time course gene expression data: evaluation of initialization and random effects. Bioinformatics. 26:370-7. 2010.

[16] D. Wu, G. K. Smyth, Camera. "A competitive gene set test accounting for inter-gene correlation". Nucleic Acids Research. 2012. 40(17).

[17] P. Boris, Hejblum , J. Skinner, R. Thiébaut. "Time-course gene set analysis for longitudinal gene expression data." Plos Computational Biology. 2015.

[18] L. H. Y. Luan. "Clustering of time-course gene expression data using a mixed-effects model with b-splines." Bioinformatics. Vol: 474-482. no. 19, 2003.

[19] C. Bécavin, N. Tchitchek , C. Mintsa-Eya, A. Lesne, A. Benecke, "Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition." Bioinformatics. 27(10):1413–1421, 2001.

[20] N. Rajicic, J. Cuschieri, D. M. Finkelstein." Identification and interpretation of longitudinal gene expression changes in trauma". Plos One. 5:E14380, 2010.

[21] D. M. Witten, R. A. Tibshirani, "Framework for feature selection in clustering". Journal of American Statistic Association. 105:713-26, 2010.

[22] P. Ma, C. I. Castillo-Davis, W. Zhong, "A data-driven clustering method for time course gene expression data". Nucleic Acids Research. 34:1261-9, 2006.

[23] K. Wang, S.K. Ng, G.J McLachlan. "Clustering of time-course gene expression profiles using normal mixture models with autoregressive random effects." BMC Bioinformatics vol 13(300) :1471-2105, 2012.