Reliability analysis of microarray data using fuzzy c-means and normal mixture modeling based classification methods

Musa H. Asyali1,2 and Musa Alci2

1Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Center, PO Box 3354, MBC-03, Riyadh 11211, Saudi Arabia and 2Department of Electrical and Electronics Engineering, Ege University, Bornova, Izmir 35100, Turkey

ABSTRACT

Motivation: A serious limitation in microarray analysis is the unreliability of the data generated from low signal intensities. Such data may produce erroneous gene expression ratios and cause unnecessary validation or post-analysis follow-up tasks. Therefore, the elimination of unreliable signal intensities will enhance reproducibility and reliability of gene expression ratios produced from microarray data. In this study, we applied fuzzy c-means (FCM) and normal mixture modeling (NMM) based classification methods to separate microarray data into reliable and unreliable signal intensity populations.

Results: We compared the results of FCM classification with those of classification based on NMM. Both approaches were validated against reference sets of biological data consisting of only true positives and true negatives. We observed that both methods performed equally well in terms of sensitivity and specificity. Although a comparison of the computation times indicated that the fuzzy approach is computationally more efficient, other considerations support the use of NMM for the reliability analysis of microarray data.

Availability: The classification approaches described in this paper and sample microarray data are available as Matlab® (The MathWorks Inc., Natick, MA) programs (mfiles) and text files, respectively, at http://rc.kfshrc.edu.sa/bssc/staff/MusaAsyali/Downloads.asp. The programs can be run/tested on many different computer platforms where Matlab is available.

Contact: asyali@kfshrc.edu.sa

1 INTRODUCTION

DNA microarray technology is a powerful and efficient means of measuring relative gene activity or expression in a variety of applications. A comprehensive review of the biological and technical aspects of microarray technology can be found in Nyugen et al. (2002) and Golub et al. (1999). In any microarray hybridization experiment, only a small fraction of the genes become expressed as a result of the investigated conditions. Thus, a large portion of microarray data comprises low signal intensities that cause variability or impair reproducibility of the measured ratios between control and experimental samples. There are also other situations that give rise to low signal values, such as the deposition of suboptimal amounts of the probes, quality of the probes or incorrect segmentation of the spots. The identification of reliable and unreliable data points before generating the gene expression ratios provides the biologist with an extra layer of protection against the false positives. In a recent study, Asyali et al. (2004) described a classification method based on univariate and bivariate normal mixture modeling (NMM) (McLachlan and Gordon, 1989; Symons, 1981; Wolfe, 1970; Duda et al., 2000; Martinez and Martinez, 2001) for the reliability analysis of microarray data. First, the Expectation Maximization (EM) algorithm (Dempster et al., 1977; Redner and Walker, 1984; Moon, 1996) was utilized to estimate the parameters of the mixture model and the class posterior probabilities. Subsequently, the Bayesian decision theory (Duda et al., 2000) was applied to find the optimal decision boundary that discriminates between the reliable and the unreliable (low) signal intensity populations, based on the estimated class posterior probabilities.

The fuzzy c-means (FCM) classification has been successfully applied to the clustering analysis of microarray hybridization data for identifying biologically relevant groups of genes (Dembele and Kastner, 2003); however, the use and efficacy of this technique for the purpose of reliability analysis of microarray data has not yet been evaluated. In this study, as an alternative to the classification based on NMM, we proposed the use of FCM classification (Bezdek, 1981; Bezdek et al., 1987; Jang et al., 1997; Wang, 1997; Ross, 1995), which is a non-parametric approach that has found widespread biomedical applications recently (Hall et al., 1992; Karlik et al., 2003; Akay, 2000), and compared the results of both approaches against the reference (or ground truth) sets that have been constructed using our experimental data. We also evaluated the overall agreement between the results of two approaches and compared their execution times on our experimental data with a publicly available large dataset.

2 SYSTEM AND METHODS

2.1 Experimental data

We used data from three independent experiments of microarray gene expression from the same cell system (monocytic leukemia cell line, THP-1, induced by the endotoxin, LPS) (Suzuki et al., 2000; Murayama et al., 1997) in order to test and compare different classification approaches. We used complementary
DNA (cDNA) microarray, which contained about 2000 cDNA distinct probes and a total of about 4000 elements (Frevel et al., 2003). The details of microarray preparation, image acquisition and intensity extraction procedures can be found in Asyali et al. (2004). Our data consist of Cy3 (green) and Cy5 (red) channel fluorescence signal intensities. After background-subtraction and normalization, both channels were natural log-transformed, as commonly performed in microarray data analysis. In our case, the log-transform also brings the distribution of the data closer to normality, which helps fitting normal mixture models. In addition to our experimental data, a publicly available dataset from a recent study (Chang et al., 2004) was also used. The microarray procedures of the study involved about 40,000 elements, representing about 36,000 different genes. We downloaded the raw Excel data file No. 17368 that corresponds to the profiling of asynchronous arm fibroblasts versus common fibroblasts from the website (http://genome-www5.stanford.edu).

Table 1 shows summary statistics, including mean, SD, median, and the correlation between the channels ($\rho_{Cy3,Cy5}$) and the number of samples ($n$), for the two channel data in the four datasets.

<table>
<thead>
<tr>
<th></th>
<th>Dataset 1</th>
<th>Dataset 2</th>
<th>Dataset 3</th>
<th>Dataset 4(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cy3</td>
<td>Cy5</td>
<td>Cy3</td>
<td>Cy5</td>
</tr>
<tr>
<td>$N$</td>
<td>3027 (53)</td>
<td>3040 (54)</td>
<td>6455 (43)</td>
<td>39168</td>
</tr>
<tr>
<td>Mean ± SD</td>
<td>6.28 ± 1.08</td>
<td>6.16 ± 1.16</td>
<td>6.21 ± 1.05</td>
<td>6.08 ± 1.15</td>
</tr>
<tr>
<td>Median</td>
<td>5.99</td>
<td>5.81</td>
<td>5.95</td>
<td>5.77</td>
</tr>
<tr>
<td>$\rho_{Cy3,Cy5}$</td>
<td>0.9255</td>
<td>0.9239</td>
<td>0.97</td>
<td>0.8672</td>
</tr>
</tbody>
</table>

The numbers in parentheses indicate the number of elements in the reference sets.

\(^a\)Dataset 4 is obtained from the study of Chang et al. (2004).

Table 1. Summary statistics for the four microarray datasets

2.2 FCM classification

Cluster analysis (Duda et al., 2000; Ross, 1995) is based on partitioning a collection of data points into a number of subgroups, where the objects in a particular group or cluster show a certain degree of closeness or similarity. (If the number of clusters is known a priori, as in our case, clustering problem turns into a classification problem, we therefore use the terms 'clustering' and 'classification' interchangeably.) The similarity measure is generally taken as the Euclidean distance between the data points. Hard clustering, also known as k-Means, assigns each data point to one and only one of the clusters, therefore the degree of membership for each data point to a particular class is either 0 or 1.

There are several applications in which the clusters have no clear or well-defined boundaries (Dembele and Kastner, 2003; Hall et al., 1992; Karlik et al., 2003). In fuzzy clustering, each data point may belong to any class with a certain possibility or ‘degree of membership’, a value between 0 and 1. As it will be noted shortly, this concept is similar to the posterior probability in the case of mixture models. The rationale behind the fuzzy clustering lies in the reality that an object or data point could be assigned to different classes. That is, if an object does not clearly fit into any of the clusters, this knowledge, expressed by the degree of membership, can be captured.

The FCM algorithm was first proposed by Bezdek (1981) and is briefed here for convenience. Below, $\epsilon$ (the number of clusters) is 2, $l = 1,2$ is the class, $k = 1,2,\ldots,n$ is the data point and $l = 1,2,\ldots,L$ is the iteration index. The norm operator $\parallel \parallel$ refers to the Euclidean norm for vectors and Frobenius norm for matrices. Following the common practice (Dembele and Kastner, 2003; Jang et al., 1997; Ross, 1995), we selected the exponent parameter $m$ (must be $>1$, also known as the fuzziness parameter) as 2: the maximum number of iterations ($L$) and termination criterion ($\epsilon$) were taken as 100 and $10^{-5}$, respectively.

Step 1. For a given dataset $X = \{x_1, x_2, \ldots, x_n\}, x_k \in \mathbb{R}^2, set \ l = 1$ and initialize $n \times 2$ partition or membership matrix $U^{(0)}$ with elements $u_{ki}^{(0)}$, such that $0 \leq u_{ki}^{(0)} \leq 1, \sum_{k=1}^{n} u_{ki}^{(0)} = 1, \forall k$. (We initialized $U$ with random numbers, normalized to make row sums equal to 1.)

Step 2. Compute $c$ mean vectors (fuzzy centroids) $v_i^{(l)}$’s as follows:

$$v_i^{(l)} = \frac{\sum_{k=1}^{n} u_{ki}^{(l)} x_k}{\sum_{k=1}^{n} u_{ki}^{(l)}}$$

Step 3. Compute the degree of membership of all data points for all clusters and update the partition matrix, i.e. obtain $U^{(l+1)}$, as follows:

$$u_{ki}^{(l+1)} = \left[ \frac{1}{\left( \sum_{j=1}^{c} \left( \| x_k - v_j^{(l)} \| / \| x_k - v_i^{(l)} \| \right)^{2/(m-1)} \right)^{1/m} } \right]$$

Step 4. Check for convergence: stop, if $\| U^{(l+1)} - U^{(l)} \| < \epsilon \ or \ l = L$, otherwise, set $l \leftarrow l+1$ and go to Step 2. The FCM algorithm converges into a solution usually rather rapidly and there is a guaranteed convergence in a finite number iterations (Bezdek et al., 1987); however, the algorithm may converge into a local minimum as well. Since algorithm runs relatively fast, it is possible to run it with several different initial conditions to check for the optimality of the resulting clustering.
M.H. Asyali and M. Alci

Fig. 1. FCM and NMM based classification results. The ‘high’ and ‘low’ refer to the classes of reliable and unreliable data points, respectively. (A–D) Correspond to the datasets 1, 2, 3 and 4. The data points classified by FCM as low and high are marked with light-gray circles and dark-gray plus marks, respectively. The FCM decision boundary is a line that passes through the points where circles and plus marks are touching. A sample FCM decision boundary, obtained by visual inspection, is shown in (B). The dashed ellipsoid lines or contours represent the components of the bivariate normal mixture model, i.e. $N(x; \mu_1, \Sigma_1)$ and $N(x; \mu_2, \Sigma_2)$. The contour correspond to the level at which the pdf of the bivariate normal density drops to 60.65% of its peak value. The peak values are attained at the centers (indicated by large black dots) of the pdfs. In other words, the contour is obtained by cutting the two-dimensional pdf at 1 SD away from the center in each direction. The dotted ellipsoid is the NMM decision boundary, obtained by equating the two weighted density components (i.e. the decision boundary is the collection points $x$ in two-dimension, for which $w_1 N(x; \mu_1, \Sigma_1) = w_2 N(x; \mu_2, \Sigma_2)$. In (D), to underline the discrepancy between FCM and NMM classification results, the areas for which there is a disagreement are annotated.

2.3 Classification using NMM
Mixture modeling is a widely used technique for probability density function (pdf) estimation (Wolfe, 1970; Martinez and Martinez, 2001) and found significant applications in various biological problems (McLachlan et al., 2002; McManus, 1983; Shoukri and McLachlan, 1994; McLachlan and Gordon, 1989). We modeled the pdf of microarray data with two bivariate normal pdfs as follows:

$$f(x) = w_1 N(x; \mu_1, \Sigma_1) + w_2 N(x; \mu_2, \Sigma_2),$$

where, $N(x; \mu_i, \Sigma_i) = (2\pi)^{-1} \det(\Sigma_i)^{-1/2} \exp\left[-(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)/2\right], i = 1, 2$ is a bivariate normal pdf with mean $\mu_i \in R^2$ and $2 \times 2$ covariance matrix $\Sigma_i$. The $w_i (\geq 0)$ denotes the weight of $N(x; \mu_i, \Sigma_i)$. For
Reliability analysis of microarray data

Table 2. The results of NMM for the four datasets (Comp. component)

<table>
<thead>
<tr>
<th>Dataset 1</th>
<th>Dataset 2</th>
<th>Dataset 3</th>
<th>Dataset 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp. 1</td>
<td>Comp. 2</td>
<td>Comp. 1</td>
<td>Comp. 2</td>
</tr>
<tr>
<td>Mean ($\mu^T$)</td>
<td>5.83</td>
<td>5.66</td>
<td>7.53</td>
</tr>
<tr>
<td>Covariance Matrix ($\Sigma$)</td>
<td>[0.23 0.17]</td>
<td>[1.64 1.52]</td>
<td>[0.02 0.20]</td>
</tr>
<tr>
<td>Weight ($\pi$)</td>
<td>0.736</td>
<td>0.264</td>
<td>0.755</td>
</tr>
</tbody>
</table>

Table 3. Comparison of the FCM and NMM classification results on the reference sets

<table>
<thead>
<tr>
<th>Dataset 1</th>
<th>Dataset 2</th>
<th>Dataset 3</th>
<th>Dataset 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive</td>
<td>26</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>True negative</td>
<td>0</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

We used EM algorithm (Dempster et al., 1977) to estimate the mixture parameters. Following the common practice, we started the algorithm with an initial estimate of the parameters obtained from k-Means algorithm (Duda et al., 2000; Martinez and Martinez, 2001; Jang et al., 1997) and iterated the Expectation and Maximization steps until the changes in the parameters were less than a small preset tolerance (0.0001) or a certain number of iterations (300) was reached.

Similar to FCM, depending on the initial conditions, the EM algorithm may also converge into a local solution, i.e. to a local maximum of the likelihood function. The EM algorithm can be run multiple times starting with different initial guesses; however, this heuristic approach is computationally costly. Fortunately, when initialized by the k-Means algorithm, the EM algorithm will always find a good or acceptable local maximum (McLachlan and Basford, 1989) that is often considered sufficient in practical applications.

3 RESULTS

We performed all the computations of FCM and NMM classifications using our in-house programs, developed under Matlab™ (The MathWorks Inc., Natick, MA), on a personal computer with 1.5 MHz Pentium IV processor and 384 MB of memory, running under Windows-2000™ operating system. Table 2 presents the NMM modeling results, i.e. mean vectors, covariance matrices and the weights, for the four datasets.

The classification performance of both the approaches against the reference sets for the first three cases, corresponding to our experimental data, are reported in Table 3. For the fourth dataset, obtained from the study of Chang et al. (2004), we do not have a reference set. The performance of the two approaches are compared by the 2 × 2 tables showing the agreement between the true state of the nature, i.e. true positive (reliable) and the negative (unreliable) data points in the reference sets, and the classification results, i.e. reliable and unreliable decisions obtained for those data points, and the

647
corresponding sensitivity and specificity rates. For datasets 1 and 2, both methods correctly classify all the true positives and true negatives, signifying sensitivity and specificity rates of 100%. However, for the third dataset, the FCM incorrectly classifies two true positives as low or unreliable (sensitivity ~93%, specificity 100%), while NMM incorrectly classifies one true negative as reliable (sensitivity 100%, specificity ~93%).

We also explored the overall agreement between the FCM and NMM classification results, i.e. the agreement between unreliable (low) or reliable (high) decisions made for all the data points in the sets. The 2×2 comparison tables along with the corresponding agreement rates are presented in Table 4. For datasets 1 and 2, the overall agreement rate between the FCM and NMM is ~95%, whereas for datasets 3 and 4, it is only ~90%.

The computation time required for executing the algorithms for all the four cases are given in Table 5. We observe that the FCM consistently takes less time to run.

Table 5. Comparison of the execution times of the FCM and NMM classification algorithms

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Execution time (s)</th>
<th>FCM</th>
<th>NMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset 1</td>
<td>0.203</td>
<td>0.750</td>
<td></td>
</tr>
<tr>
<td>Dataset 2</td>
<td>0.234</td>
<td>0.765</td>
<td></td>
</tr>
<tr>
<td>Dataset 3</td>
<td>0.406</td>
<td>2.625</td>
<td></td>
</tr>
<tr>
<td>Dataset 4</td>
<td>2.985</td>
<td>7.797</td>
<td></td>
</tr>
</tbody>
</table>


discussion and conclusion

The microarray technology lets the biologist study the expression or the activity of thousands of genes at the same time but only a fraction of genes are differentially expressed and low signal intensities constitute a relatively large portion of the data. Such low signal intensities may give rise to erroneous gene expression ratios or false positives. Therefore, careful filtering of such signals before the subsequent steps of the analysis is essential. Various techniques (Brody et al., 2002; Hughes et al., 2000; Bilban et al., 2002; Tran et al., 2002; Fielden et al., 2002) to study the microarray spot accuracy and identify the true array signals have been suggested in the literature. Recently, Asyali et al. (2004) suggested a NMM-based approach and successfully demonstrated its advantages over the existing techniques. The major novelty of their approach was the assignment of ‘reliability probability’ to the raw data points. In this study, we have explored the possibility of accomplishing the same signal classification goal, i.e. reliable versus unreliable, using the popular FCM classification technique.

The FCM assigns a ‘degree of membership’ to each data point as well, similar to the ‘posterior probability’ assignment of the NMM. This feature is very important because depending on the characteristic of the data, the biologist may want to change the default cutoffs to make the ‘reliable’ or ‘unreliable’ calls for the data points. For instance, we assumed that if the degree of membership (FCM) or the posterior probability (NMM) for a data point for belonging to the low or unreliable class is ~0.5, then the point should be identified as reliable. However, suppose this type of reliability analysis based filtering of microarray data turn out to be too restrictive. Then, one can relax the reliability constraint slightly and decide for reliable class if the degree of membership (FCM) or the posterior probability...
(NMM) for a data point belonging to the reliable class is >0.45. In any case, the estimated 'degree of membership' or 'posterior probability' can be kept in perspective to assess the reliability of the gene expression ratios. Essentially, making this type of ‘hard’ calls or filtering is not even necessary, as the basic idea is to have a ‘degree of reliability’ or ‘probability of reliability’ assigned to each data point so that one can know the reliability of corresponding gene expression ratios.

Therefore, we thought that a comparison of FCM and NMM classification approaches, which both seem to be suitable for the reliability analysis of microarray data, would be interesting to do. To this end, we have applied both algorithms on four datasets and assessed their performance by checking the classification decisions, ‘reliable’ versus ‘unreliable’, against the information in the reference sets where available (Table 3) and also by comparing the overall agreement between the results of the two approaches (Table 4).

Based on the performance comparison against the reference sets, which indicates that both algorithms are performing equally well (Table 3), and considering the speed advantage of the FCM (Table 5), one may jump to the conclusion that it is advantageous to use FCM. Especially in the case of batch processing of large datasets, the speed advantage of FCM may be an appealing factor. However, a closer look into the classification results, particularly for the third and fourth datasets shown in Figure 1C and D reveals that the decision boundary (and corresponding decision region), which is identified by the NMM has some unique properties. Both FCM and NMM decision regions for the unreliable data lie in the lower left quarter of the two-dimensional data space, which is quite sensible, as in our context ‘unreliability’ is directly related with the ‘lowness’ of signal values. On the other hand, the decision boundary of the NMM is aligned with the Cy3 = Cy5 axis. This means that when both Cy3 and Cy5 channel signals are ‘low’ and ‘unbalanced’ the NMM will most likely identify those points as reliable, whereas the FCM will fail to do so. This point is clearly seen in Figure 1D, the regions annotated as ‘FCM Low, NMM High’ most probably correspond to reliable data points. For the region marked as ‘FCM Low, NMM High’, one may argue that the FCM is reaching a more fair decision than NMM, as the NMM decision boundary reaches too deep into the region of bivariate normal density component with the higher mean. (The NMM decision boundary is almost touching the center of the component with the higher mean. This is quite possible, depending on the parameter setting of the density components and the class prior probabilities, i.e. weights.) However, for these points, the gene expression ratio, i.e. Cy5/Cy3 ratio, will not be interesting anyway (a ratio close to 1 does not signify any differential gene expression). This observation, i.e. the alignment of the decision boundary of the NMM along the identity line, led us to conclude that NMM is superior to FCM in terms of identifying or assessing the reliability of microarray data.

REFERENCES


