A Combined Markov and Noise Clustering Modeling Method for Cell Phase Classification

DAT T. TRAN
School of Information Sciences and Engineering
University of Canberra
Canberra, ACT 2601
AUSTRALIA
dat.tran@canberra.edu.au

TUAN D. PHAM
Bioinformatics Applications Research Centre
School of Information Technology
James Cook University
Townsville, QLD 4811
AUSTRALIA
tuan.pham@jcu.edu.au

Abstract: This paper proposes a classification method of cell nuclei in different mitotic phases using a combined Markov and noise clustering modeling technique. The method was tested with the data set containing 379519 cells in 892 cell sequences for 5 phases extracted from real image sequences recorded at every fifteen minutes with a time-lapse fluorescence microscopy. Experimental results showed that the proposed method performed better than the *k*-means modeling method.

Keywords: Cell phase classification, noise clustering, Markov modeling, *k*-means clustering, *n*-nearest neighbor method.

1 Introduction

Stages of an automated cellular imaging analysis consist of segmentation, feature extraction, classification, and tracking of individual cells in a dynamic cellular population. Automatic classification of cell phases is the most difficult task of such analysis [2]. The increasing quantity and complexity of image data from dynamic microscopy renders manual analysis unreasonably time-consuming. Therefore, automatic techniques for analyzing cell-cycle progress are of considerable interest in the drug discovery process.

We applied computational techniques for classifying individual cell phase changes during a period of time. To extract useful features for the cell-phase classification task, the image segmentation of large image sequences acquired by time-lapse microscopy is necessary. The extracted data can then be used to analyze cell phase changes under drug influence. Segmenting nuclei in time-lapse microscope can be performed by various methods such as thresholding, region growing, or edge detection [2]. Most of these algorithms take into account either the morphological information or the intensity information of the image. Problems may arise when trying to segment touching nuclei because it is very difficult to define the boundary of each individual nuclear. Watershed techniques can be used to segment touching objects [1]. To deal with the over-segmentation problem a post process is needed to merge the fragments. Umesh and Chaudhuri [15] used a connectivity based merging method to merge a tiny cell fragment with a nearby cell if it shares the maximum boundary with that cell. This method can only merge small fragments and fails if the size of cell fragments is above a preset value. The bigger fragments are considered as cell by this method. Bleau and Leon [1] used an iterative trial and test approach to merge small regions with their nearby larger regions based on a set of volume, depth, and surface criteria. These authors applied their method to segment the vesicles in live cells; however no experimental results were reported.

To automate the process of classifying cellular phases using time-lapse fluorescence microscopic image sequences, we first apply a shape-and-size

based method which merges the over-segmented nuclear fragments. Secondly we extract useful features to discriminate the shapes and intensities of different image cell phases. We then use these image features to train noise clustering phase model and Markov models. These models are then combined to classify unknown cancer cells at different phases. We also compare the combined modeling method with the popular *k*-means modeling method. Experimental results showed that the proposed method performed better than the *k*-means modeling method.

The paper is organized as follows. Section 1 presents a brief introduction of automated cellular imaging analysis and current methods. Section 2 presents the modeling method using *k*-means clustering. The proposed combined Markov and noise clustering modeling method is presented in Sections 3, 4 and 5. Experimental results are presented in Section 6. Finally we conclude the paper in Section 7.

2 *k*-Means Modeling

Given a training set of T feature vectors $X=\{x_1,x_2,...,x_T\}$, where each source vector $x_t=(x_{t1},x_{t2},...,x_{tK})$ is of K dimensions. Let $U=\{u_{nt}\}$ be a matrix whose elements are memberships of x_t in the n-th cluster, n = 1,..., N and t = 1,..., T. The k-partition space for X is the set of matrices U such that

$$u_{nt}\in\{0,1\},\ \sum_{n=1}^{N}u_{nt}=1 \text{ and } 0<\sum_{t=1}^{T}u_{nt}<T \quad (1)$$

where $u_{nt}=u_n(x_t)$ is 1 or 0, according to whether x_t is or is not in the n-th cluster, $\sum_{n=1}^{N}u_{nt}=1,\forall t$ means each x_t is in exactly one of the N clusters, and $0<\sum_{t=1}^{T}u_{nt}<T,\forall n$ means that no cluster is empty and no cluster is all of X because of $1<N<T$.

The k-means modeling technique is based on minimization of the sum-of-squared-errors function as follows

$$J(U,\lambda;X)=\sum_{n=1}^{N}\sum_{t=1}^{T}u_{nt}d_{nt}^2 \quad (2)$$

where $U=\{u_{nt}\}$ is a hard k-partition of X, $\lambda=\{c_1,c_2,...,c_N\}$ is the set of N cluster centers $c_n=(c_{n1},c_{n2},...,c_{nK})$, n = 1,..., N, and $d_{nt}=\|x_t-c_n\|_2$, where $\|e_t\|_2$ is the L_2 norm or Euclidean norm of the vector e_t and defined as

$$\|e_t\|_2=\sqrt{e_{t1}^2+e_{t2}^2+...+e_{tK}^2} \quad (3)$$

The k-means modeling algorithm is summarized as follows

1. Given a training set $X=\{x_1,x_2,...,x_T\}$, where $x_t=(x_{t1},x_{t2},...,x_{tK})$, t = 1,..., T.
2. Initialize membership values u_{nt}, n = 1,..., N and t = 1,..., T, at random.
3. Given $\varepsilon>0$ (small real number)
4. Set i = 0 and $D^{(i)}=0$. Iteration
   a. Compute cluster centers

$$c_n=\sum_{t=1}^{T}u_{nt}x_t\Big/\sum_{t=1}^{T}u_{nt},\ n=1,...,N \quad (4)$$

   b. Compute d_{nt} and $D^{(i+1)}$

$$d_{nt}=\|x_t-c_n\|_2 \quad (5)$$

$$D^{(i+1)}=J(U,\lambda;X) \quad (6)$$

   c. Update membership values

$$u_{nt}=\begin{cases}1: d_{nt}<d_{jt}, j=1,...,N, j\neq n\\ 0: otherwise\end{cases} \quad (7)$$

5. If

$$\frac{|D^{(i+1)}-D^{(i)}|}{D^{(i+1)}}>\varepsilon \quad (8)$$

then set $D^{(i)}=D^{(i+1)}$, $i=i+1$ and go to step *a*.

3 Noise Clustering Modeling

Most of clustering methods have a disadvantage in the problem of sensitivity to outliers. An idea of a noise cluster has been proposed [3] to deal with noisy data or outliers for fuzzy clustering methods.

The noise is considered to be a separate class and is represented by a prototype that has a constant distance δ from all feature vectors. Therefore the sum of memberships for the good clusters should be smaller than one. This allows noisy data and outliers to have arbitrarily small membership values in the good clusters.

Given a training set of T feature vectors $X = \{x_1, x_2, ..., x_T\}$, where each source vector $x_t = (x_{t1}, x_{t2}, ..., x_{tK})$ is of K dimensions. Let $U = \{u_{nt}\}$ be a matrix whose elements are fuzzy memberships of x_t in the n-th cluster, $n = 1, ..., N$ and $t = 1, ..., T$.

The fuzzy partition space for X is the set of matrices U such that

$$0 \le u_{nt} \le 1, \quad \sum_{n=1}^{N} u_{nt} = 1 \text{ and } 0 < \sum_{t=1}^{T} u_{nt} < T \tag{9}$$

where $0 \le u_{nt} \le 1$ means it is possible for each x_t to have an arbitrary distribution of membership among the N fuzzy clusters.

The noise clustering technique is based on minimization of the fuzzy sum-of-squared-errors function as follows [3]

$$J(U, \lambda; X) = \sum_{n=1}^{N}\sum_{t=1}^{T} u_{nt}^m d_{nt}^2 + \sum_{t=1}^{T} \delta^2 (1 - \sum_{n=1}^{N} u_{it})^m \tag{10}$$

where $U = \{u_{nt}\}$ is a fuzzy partition of X, $\lambda = \{c_1, c_2, ..., c_N\}$ is the set of N cluster centers $c_n = (c_{n1}, c_{n2}, ..., c_{nK})$, $n = 1, ..., N$, $m > 1$ denotes the degree of fuzziness, δ is constant distance and $d_{nt} = \| x_t - c_n \|_2$.

The noise clustering modeling algorithm is summarized as follows

1. Given a training set $X = \{x_1, x_2, ..., x_T\}$, where $x_t = (x_{t1}, x_{t2}, ..., x_{tK})$, $t = 1, ..., T$.
2. Initialize fuzzy membership values u_{nt}, $n = 1, ..., N$ and $t = 1, ..., T$, at random.
3. Given $\varepsilon > 0$ (small real number)
4. Set $i = 0$ and $D^{(i)} = 0$. Iteration
   a. Compute cluster centers

$$c_n = \sum_{t=1}^{T} u_{nt}^m x_t \Big/ \sum_{t=1}^{T} u_{nt}^m, \quad n = 1, ..., N \tag{11}$$

   b. Compute d_{nt} and $D^{(i+1)}$

$$d_{nt} = \| x_t - c_n \|_2 \tag{12}$$

$$D^{(i+1)} = J(U, \lambda; X) \tag{13}$$

   c. Update membership values

$$u_{nt} = \frac{1}{\sum_{i=1}^{N} (d_{nt}/d_{it})^{2/m-1} + (d_{nt}/\delta)^{2/m-1}} \tag{14}$$

5. If

$$\frac{| D^{(i+1)} - D^{(i)} |}{D^{(i+1)}} > \varepsilon \tag{15}$$

then set $D^{(i)} = D^{(i+1)}$, $i = i + 1$ and go to step *a*.

4 Markov Modeling

Let $X = \{X^{(1)}, X^{(2)}, ..., X^{(L)}\}$ be a set of L cell sequences, where $X^{(k)} = (x_1^{(k)}, x_2^{(k)}, ..., x_{T_k}^{(k)})$ is a cell sequence of T_k cells, $k = 1, 2, ..., L$ and $T_k > 0$. Let $V = \{v_1, v_2, ..., v_M\}$ be the set of M cell phases regarded as M states in a Markov chain. Define the following parameters

$$q = [q(i)], \qquad q(i) = P(x_1^{(k)} = v_i) \tag{16}$$

$$p = [p(i,j)], \quad p(i,j) = P(x_t^{(k)} = v_j \mid x_{t-1}^{(k)} = v_i) \tag{17}$$

where $k = 1, ..., L$, $i, j = 1, ..., M$, M is the number of cell phases. The set $\lambda = (q, p)$ is called a Markov cell phase model that represents sequences of phases observed in the set of L cell sequences. A method to calculate the model set $\lambda = (q, p)$ is presented as follows.

The Markov model λ is built to represent the sequence of states $V = (v_1, v_2, ..., v_T)$, where $T = \sum_{k=1}^{L} T_k$, therefore we should find λ such that the following probability $P(X = V \mid \lambda)$ is maximized

$$P(\mathrm{X}=\mathrm{V}\mid\lambda)=\prod_{i=1}^{M}[q(i)]^{n_i}\prod_{j=1}^{M}[p(i,j)]^{n_{ij}} \qquad (18)$$

where n_i denotes the number of values $x_1^{(k)}=v_i$ and n_{ij} denotes the number of pairs $(x_{t-1}^{(k)}=v_i, x_t^{(k)}=v_j)$ observed in the sequence $X^{(k)}$. The probability in (20) can be rewritten as follows

$$\log[P(\mathrm{X}=\mathrm{V}\mid\lambda)]=\sum_{i=1}^{M}n_i\log q(i)+\sum_{i=1}^{M}\sum_{j=1}^{M}n_{ij}\log p(i,j) \qquad (19)$$

Since $\sum_{i=1}^{M}q(i)=1$ and $\sum_{j=1}^{M}p(i,j)=1$, the Lagrangian method is applied to maximize the probability in (19) over λ. We have

$$q(i)=\frac{n_i}{\sum_{s=1}^{M}n_s} \qquad p(i,j)=\frac{n_{ij}}{\sum_{s=1}^{M}n_{is}} \qquad (20)$$

The equations in (20) are used to determine the Markov cell phase models from the training set.

5 Markov-Noise Clustering Modeling

The noise clustering and Markov modeling methods can be combined to build better cell phase models. The training and classification procedures of this combined Markov and noise clustering modeling method are summarized as follows

Training:

1. Given **X** as the training set of all cell phases.
2. Train M phase models as follows:
   - Divide the set **X** into M distinct subsets $X^1, X^2, \ldots, X^M$, where each X^i contains only cells of phase i.
   - For each subset X^i, train a noise clustering phase model using the algorithm in Section 3.
3. Train a Markov model for all phases as follows:
   - Align cells in the set **X** as sequences of cells
   - Extract $X^{(1)}, X^{(2)}, \ldots, X^{(L)}$ as L phase sequences from the set **X**
   - Using L phase sequences, compute the Markov model using the equations in (20).

Classification:

1. Given $X=(x_1, x_2, \ldots, x_T)$ as an unknown cell sequence. The task is to classify phase for each cell in the sequence.
2. Classify phase for the first cell x_1 in the sequence as follows:
   - Compute the M distances $d_i=\| x_1-c^{(i)}\|_2$, $i=1, 2, \ldots, M$ between the unknown cell x_1 and the closest cluster center $c^{(i)}$ in the i-th noise clustering phase model.
   - Compute the similarity score $S(x_1, \lambda_i)$

$$S(x_1,\lambda_i)=\frac{q(i)}{\sum_{k=1}^{M}(d_i/d_k)^{1/(m-1)}} \qquad (21)$$

   where $m>1$
   - Assign x_1 to the phase i^* that has the maximum score

$$i^*=\arg\max_i S(x_1,\lambda_i) \qquad (22)$$

3. For each cell x_t, $t=2, 3, \ldots, T$, classify it as follows
   - Compute the M distances $d_i=\| x_t-c^{(i)}\|_2$, $i=1, 2, \ldots, M$ between the unknown cell x_t and the closest cluster center $c^{(i)}$ in the i-th noise clustering phase model.
   - Calculate the similarity score $S(x_t, \lambda_i)$

$$S(x_t,\lambda_i)=\frac{p(i^*,i)}{\sum_{k=1}^{M}(d_i/d_k)^{1/(m-1)}} \qquad (23)$$

   where $m>1$ and i^* is the classified phase of the previous cell
   - Assign x_t to the phase i^* that has the maximum score.

$$i^*=\arg\max_i S(x_t,\lambda_i) \qquad (24)$$

6 Experimental Results

Nuclear sequences were provided by the Department of Cell Biology at the Harvard Medical School. The data set consists of 892 cell sequences labeled from 1 to 892. These sequences have different lengths, ranging from 18 to 482 cells. To classify the shape and intensity differences between different cell phases, a set of 7 features is extracted. These features include maximum intensity, mean, stand deviation, major axis, minor axis, perimeter, and compactness. There are 5 phases to be classified: interphase, prophase, metaphase, anaphase, and arrested metaphase.

Because the feature values have different ranges, the scaling of features was therefore necessary by calculating the z-scores [2]

$$z_{ij} = \frac{x_{ij} - m_j}{s_j} \tag{24}$$

where x_{ij} is the j-th feature of the i-th sequences, m_j the mean value of all T cells for feature j, and s_j the mean absolute deviation, that is

$$s_j = \frac{1}{T}\sum_{i=1}^{T} | x_{ij} - m_j | \tag{25}$$

We then divided the data set into 5 subsets for training 5 noise clustering phase models and a subset for classification. Each of the 5 training sets for 5 phases contained 5000 cells, which were extracted from the cell sequences labeled from 590 to 892. These sequences were also used to calculate the Markov model. The classification set contained sequences labeled from 1 to 589. There were 249,547 cells in this classification set.

Figure 1 presents the experimental results for cell phase classification using the k-means modeling, noise clustering modeling and the combined Markov and noise clustering modeling method. The degree of fuzziness m was set to 1.1. For simplicity, the constant distance δ in (14) was set such that $(d_{nt}/\delta)^{2/m-1} = 2$ for all n and t. The number of clusters was set to 4, 8, 16, 32, 64 and 128, respectively. It can be seen from Figure 1 that the combined Markov and noise clustering modeling method achieved the better classification rates in all values of the number of clusters. The highest classification rate is 88.91% for the combined method with the number of clusters = 128.

Table 1 presents the confusion matrix for the combined Markov and noise clustering modeling method. The number of clusters was set to 128.

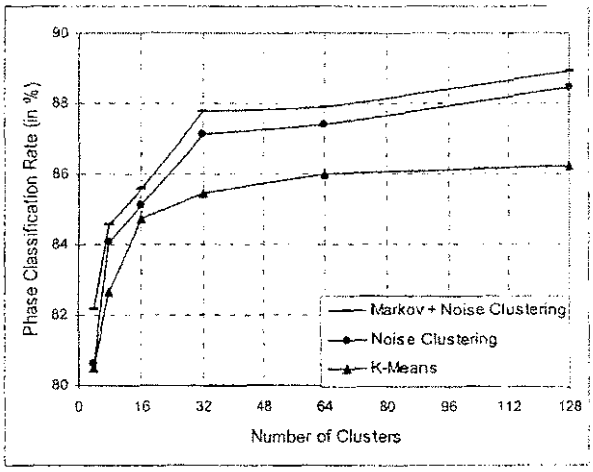

Fig. 1. Phase classification rate (in %) for k-means, noise clustering and combined Markov & noise clustering phase modeling methods.

Model size = 128	*Interphase*	*Prophase*	*Metaphase*	*Anaphase*	*Arrested Metaphase*
Interphase	**148477**	1406	5511	1163	0
Prophase	927	**24514**	2224	93	70
Metaphase	3430	5156	**15054**	1787	999
Anaphase	86	8	571	**17524**	0
Arrested Metaphase	0	1989	2192	74	**16292**

Table 1. Confusion matrix for phase classification using Combined Markov & Noise Clustering phase modeling method. Model size = 128. Total cells tested: 249,547.

7 Conclusion

We have presented the combined Markov and noise clustering modeling method for cell phase classification. Cell features were used to train the five phase models, which were interphase, prophase, metaphase, anaphase, and arrested metaphase. Phase information in cell sequences was used to train Markov model. The combined Markov and noise clustering modeling method was achieved

better classification results comparing with the noise clustering and *k*-means modeling methods.

Acknowledgements

The data set was provided by our collaborator: HCNR Center for Bioinformatics, Harvard Medical School. ARC Discovery Project 0665598 Grant entitled "An Automated Bioimaging System for High-Content Cell-Cycle Screening" provided to the second author (Tuan D. Pham) is acknowledged.

References:

[1] Bleau A., and Leon J.L., Watershed-based segmentation and region merging, *Computer Vision and Image Understanding*, Vol. 77, pp. 317-370, 2000.

[2] Chen, X., Zhou, X., and Wong, S.T.C., Automated segmentation, classification, and tracking cancer cell nuclei in time-lapse microscopy, *IEEE Trans. on Biomedical Engineering*, in press, 2005.

[3] Dave, R. N., Characterization and detection of noise in clustering, *Pattern Recognition Letters*, Vol. 12, No. 11, pp. 657-664, 1991.

[4] Duda, R.O. and Hart P.E., *Pattern classification and scene analysis*, John Wiley & Sons, New York, 1973.

[5] Dunkle, R., Role of image informatics in accelerating drug discovery and development, *Drug Discovery World*, Vol. 7, pp. 7-11, 2002.

[6] Fox, S., Accommodating cells in HTS, *Drug Discovery World*, Vol. 5, pp. 21-30, 2003.

[7] Feng, Y., Practicing cell morphology based screen, *European Pharmaceutical Review*, Vol. 7, pp. 75-82, 2002.

[8] Hiraoka, Y., and Haraguchi, T., Fluoresence imaging of mammalian living cells, *Chromosome Res*, Vol. 4, pp. 173-176, 1996.

[9] Kanda, T., Sullivan, K. F., and Wahl G. M., Histone-GFP fusion protein enables sensitive analysis of chromosome dynamics in living mammalian cells, *Current Biology*, Vol. 8, pp. 377-385, 1998.

[10] MacAulay, C., and Palcic, B. A., Comparison of some quick and simple threshold selection methods for stained cells, *Anal. Quant. Cytol. Histol*, Vol. 10, pp. 134-138, 1998.

[11] Murphy, D.B., *Fundamentals of light Microscopy and Electronic Imaging*, Wiley-Liss, 2001.

[12] Tran, D. and Wagner, M., A Proposed Fuzzy Pattern Verification System, *Proceedings of the FUZZ-IEEE 2001 Conference*, Vol. 2, pp. 932-935. 2001.

[13] Tran, D. and Wagner, M., Noise Clustering-Based Speaker Verification, *Lecture Notes in Computer Science: Advances in Soft Computing* - AFSS 2002, N.R. Pal, M. Sugeno (Eds.), pp. 325-331, Springer-Verlag, 2002.

[14] Tran, D., and Pham, T., Fuzzy and Markov models for written language verification, *Proceedings of WSEAS Conferences*, Lisbon, Portugal, June 16-18, 2005.

[15] Tran, D., Pham, T. and Zhou, X., Cell Phase Identification using Fuzzy Gaussian mixture models. *Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 465-468, 2005.

[16] Umesh, A.P.S., and Chaudhuri B.B., An efficient method based on watershed and rule-based merging for segmentation of 3-D histopathological images, *Pattern Recognition*, Vol. 34, pp. 1449-1458, 2001.

[17] Yarrow, J.C., et al., Phenotypic screening of small molecule libraries by high throughput cell imaging, *Comb Chem High Throughput Screen*, Vol. 6, pp. 279-286, 2003.

[18] Zhou, X., Chen, X., King, R., and Wong, S.T.C., Time-lapse cell cycle quantitative data analysis using Gaussian mixture models, *Life Science Data Mining*, S.T.C. Wong and C.S. Li (Eds.), World Scientific, in press.