33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

An Improved Classification Scheme for Chromosomes with Missing Data

Enea Poletti, Alfredo Ruggeri, and Enrico Grisan

Abstract— **Karyotyping, or the automatic classification of human chromosomes, is mostly based on the analysis of the chromosome specific banding pattern. Unfortunately, the most informative phases of the cell division cycle are composed of long chromosomes that easily overlap: the involved banding pattern information is corrupted, resulting in a drastic increase of the classification error.**

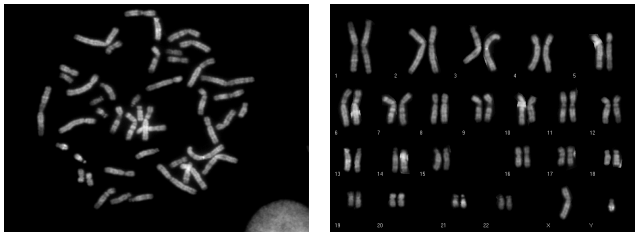
Assuming the availability of a probabilistic classifier, the improvement of the classification of chromosomes with corrupted data would require the additional estimation of the joint probability density of the observed and missing data for each chromosome class. Given the number of classes, the possible position and extension of the corrupted data within a chromosome, and the dimensionality of the feature space, a reliable estimation would need an impossible number of training samples. We chose to circumvent the estimation problem by developing a statistical generative model of the pattern of each class, so that the corrupted part can be substituted with a partial pattern synthetically generated from the model. This allows to obtain a Monte Carlo estimate of the maximum a posteriori probability for the class given the observation and the missing data, which reduces to a simple voting scheme if the *a priori* probability for each class is equal. Moreover, this Monte Carlo classification is superior to the voting scheme based on the simple imputation of the classes mean to the missing data.

I. Introduction

Karyotype analysis is an important screening and diagnostic procedure routinely performed in clinical cytogenetic labs. Chromosome are first stained with a fluorescent dye, and then imaged through a microscope for subsequent analysis and classification. Each chromosome in the image has to be identified and assigned to one of 24 classes [1] (Fig. 1).

The aim of an automatic karyotyping system is to assign each chromosome to one of the 24 possible classes, by exploiting chromosome features extracted from the image. One of the most important feature is the density profile [2], [3] that is a representation of the banding pattern of each chromosome class (Fig. 2b), and that can be obtained only after the axis of the chromosome has been estimated (Fig. 2a) [4], [2], [3].

All the previous work dealing with the classification of chromosome disregard the fact that the most informative (in terms of banding pattern) phases of the division cycle are composed of long chromosomes that easily overlap (Fig. 3a): the number of overlaps reported varies between 4% [5] and 11% [6]. Within the overlapping region, the information extracted is the superposition of the banding pattern belonging to the different chromosomes involved, and often the resulting hyper-fluorescence saturates the image intensity (Fig. 3). These effects have a twofold consequence: the banding pattern information of the two chromosome is corrupted, making the information useless or even misleading for an automatic classifier, and every procedure aiming at normalizing the intensity information describing the banding pattern will results in a suppression of the information coming from the overlap-free region (Fig. 3b). As a consequence, an effective and successful automatic karyotyping system has to tackle the problem of chromosome classification even in presence of missing or corrupted data. Generally, there



(a) (b)

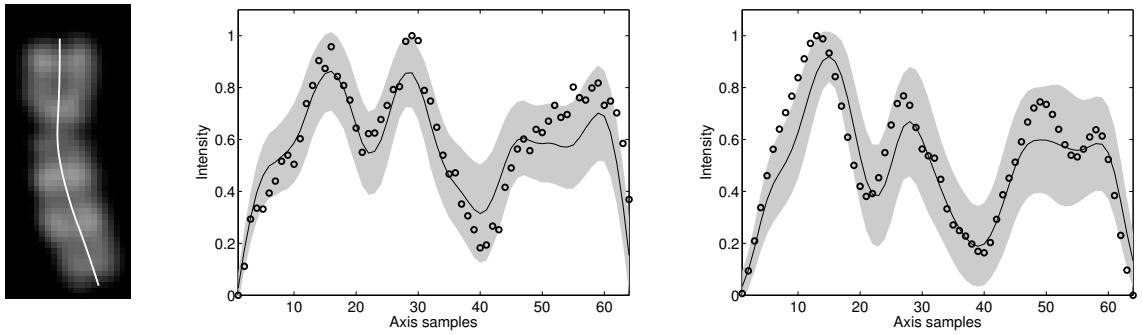
Fig. 1. Typical Q-band prometaphase image acquired with PAL resolution (a), and the manual kariotyping of the chromosomes (b)

are three ways of dealing with missing data [7], under the assumption of data Missing At Random (MAR) or Missing Completely At Random (MCAR), that is usually a reasonable assumption in chromosome banding data. The simplest way of proceeding would be to discard the missing data (*marginalization* [8]), but this approach can be used only if the amount of missing data is small. When two chromosomes overlap, the corrupted region may be as long as half the chromosome length, so that this approach is clearly unfeasible.

The second approach is to rely on the robustness of the learning algorithm, letting it to deal with missing values in the training phase. The last is to replace the missing values by estimated ones (*imputation*), either by imputing each missing value with a statistical representative of the non-missing values (e.g. through the mean or the median or by a likelihood estimation [9]), or by incorporating the classification label into the generative model [10]).

In karyotyping, successful classification schemes based on neural network and support vector machines can not be adapted to a marginalization approach, and they provide very high classification error on corrupted chromosomes. In order to tackle this problem, we propose to build a generative a model for the class banding patterns, then obtaining a

Authors are with the Department of Information Engineering, University of Padova, Via Gradenigo 6/a, 35131 Padova, Italy. enrico.grisan@dei.unipd.it



(a) (b) (c)

Fig. 2. (a) Original image of a class "7" chromosome and its axis estimation. (b) Density profile of the actual chromosomes and (c) of the generative model. Circles indicate a single density profile instance, while solid line and shaded area indicate class "7" average and standard deviation, respectively.

set of synthetic patterns by replacing the missing data with profiles obtained by the generative models for each class. The classification of this set will provide a Monte Carlo estimate of the probability a posteriori for each class. We will show that the classification performance are superior both to the direct classification of the corrupted pattern, and to the voting scheme obtained by replacing the missing data with simple statistics (the class means) on real chromosome data.

II. Materials

1) Chromosome Data: Q-band images are cytogenetic data obtained by staining the chromosomes with quinacrine, appearing with dark background, onto which the chromosomes stand out as light objects, with bright and dark banding. The dataset we used is composed of 119 cells, containing a total number of 5474 chromosomes. The images were sequentially acquired during routine laboratory analysis (Fig. 1a) and then an expert cytologist manually classified the chromosomes composing the karyotype (Fig. 1b). 342 overlapping chromosomes were identified, and each chromosome region involved in an overlap was manually outlined: this dataset is available for download at http://bioimlab.dei.unipd.it.

2) Feature Extraction: After estimating the chromosomes axis by means of the algorithm described in [5], the density profile of each chromosome is extracted. The density profile is meant to be a representation of the chromosome banding patter and is obtained as the mean intensity of the pixels along the chromosome diameter at a number of discrete sampling points. The number of sampling points is set to $M = 64$. This value was empirically determined as the best compromise between discriminating power and resiliency to noise of these feature. To reduce the intra-class feature variance and in order to make the measurements comparable among different images, the density profile has been normalized so to make its values are in the range $[0, 1]$ and an automatic polarisation step is performed, based on a binary classifier described in [11].

III. Methods

In a classical Bayesian framework, where we have K classes c_k and an observed feature vector o, the classification problem is formalized as:

$$c^* = \max_c p(c_k|o) = \frac{p(o|c_k)p(c_k)}{\sum_k p(o|c_k)} \tag{1}$$

When part of the feature vector o is missing, with the missing part being m, the problem becomes:

$$c^* = \max_c p(c_k|o) = \max_c \int_\Omega p(c_k|o,m)p(m)d\Omega \tag{2}$$

with Ω the domain of the missing vector m: the class with the maximum probability given the observation, is the one maximizing $p(c_k|o)$ considering all possible patterns of the missing values.

The joint probability density of m and o and the probability density of m can be very hard to estimate, especially when, as in the karyotyping problem, the number of classes and the dimensionality of the feature space are high, and the position and extension of the missing data within the feature vector o are variable. Hence we try to build a simple generative model for the missing data $p(m|o, c_k)$ for each class c_k, so to obtain an estimate of the integral in Eq. 2 via a Monte Carlo procedure:

$$p(c_k|o) \simeq \frac{1}{I}\sum_i^I p(c_k|o,m_i)p(m_i|o,c_k) \tag{3}$$

with $p(m_i|o, c_k)$ a synthetic pattern generated for the missing data, given the observed data and the class under consideration.

A. Density Profile Generative Model

The samples of the density profile $F(i)$, $i = 1, \ldots, M$ of each class is a smoothed and noisy version of the banding pattern: this sequence can be represented by the superposition of a mixture model of N gaussian distributions (see Fig. 2b):

$$F(i) \simeq g(i;\boldsymbol{s}) = \sum_{n=1}^N g_n(i;\boldsymbol{s}) = \sum_{n=1}^N \frac{p_n}{\sqrt{2\pi}\sigma_n} e^{-0.5\frac{(i-\mu_n)^2}{\sigma_n^2}} \tag{4}$$

with $\boldsymbol{s} = \{p_1, \ldots, p_N, \mu_1, \ldots, \mu_N, \sigma_1, \ldots, \sigma_N\}$ $3N$-dimensional parameter vector and the weighting factors p_i

normalized:

$$\sum_{n=1}^{N} p_n = 1$$

The parameter vector $\boldsymbol{s}$ describing each density profile is estimated through an *expectation-maximization*[7] procedure, that involves the iteration of two steps until convergence: the evaluation of the expected log-likelihood of the data given the current estimates of the parameters (*E-step*), and a subsequent refinement of the estimates obtained maximising the obtained log-likelihood with respect to the parameters (*M-step*). Since one of the major problem in parameter estimation is a careful set of initial estimates, we would like to provide the algorithm with an initial distribution which present very distinct and non ambiguous modes. We therefore will set a small initial variance for every distribution , and equal weighting factors, whereas the initial means μ_i are provided by a fuzzy c-means clustering with N classes: the identified cluster centers are the initial mean values.

Each chromosome can thus be identified by the set of $3N$ parameters identifying its mixture-model, with the value of modes N being a class-specific value. Given the number of data points d composing each chromosome density profile, a measure of the goodness of a model can be obtained by evaluating the unbiased Akaike Information Criterion (AICU) proposed in [12]:

$$AICu = \log\left(\frac{RSS}{d}\right) + \frac{d + 3 \cdot N}{d - 3 \cdot N - 2} \quad (5)$$

$$RSS = \sum_i ((g(i;\boldsymbol{s}) - DP(i))^2) \quad (6)$$

For each chromosome c_k of a class k, the $3N$ parameters of the mixture models are estimated, and the corresponding value of the $AICu(c_k, N)$ measure evaluated, for N spanning the range $[1, 6]$. Then, the mean value $\mu_k(N) = E_{c_k}[AICu(c_k, N)]$ for the $AICu$ measure over all chromosome of the class, for each model, is evaluated, so that the model yielding the minimum value of it is retained as the best:

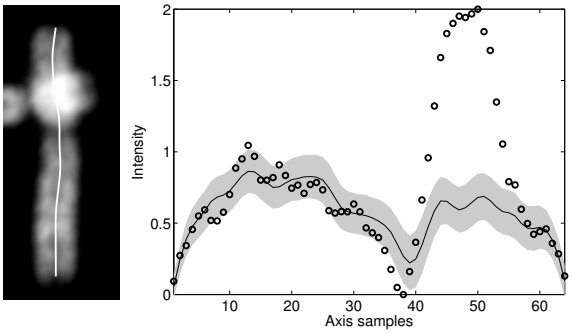
$$N_k = argmin(\mu_{AICu}(N)) \quad (7)$$

Given the gaussian mixture model description available for each chromosome, it is possible to obtain a statistical description of the parameters within each class. By assuming that the parameter distribution of each of the 24 chromosome classes can be entirely described by its first two moments, for the k^{th} class we can compute:

$$\mu(k)_s = E[\boldsymbol{s}|k] \quad (8)$$

$$\Sigma(k)_s = Var[\boldsymbol{s}|k] \quad (9)$$

With this description, given a class k, we are able to draw a set of parameters $\boldsymbol{s}_k^*$ from a $3N$-dimensional normal distribution $\mathcal{N}(\mu(k), \Sigma(k))$, and then generate a synthetic profile $g(i, \boldsymbol{s}_k^*)$ (Fig. 2c).



(a) (b)

Fig. 3. (a) Original image of an overlapping class "2" chromosome, and its axis estimation. (b) Circles indicate the chromosome density profile, while solid line and shaded area indicate class "2" average and standard deviation, respectively.

B. Classification function

Chromosome classes are not linearly separable, whereas the problem dimensionality makes a non-parametric probability density estimation, to be used in a Bayesian classifier, unfeasible with a reasonable number of training samples. Following the approach used to tackle the same problem [13], [3], [14], [15], [16], we use an Artificial Neural Network (ANN) classifier. In fact, given an input vector F, the output vector $h = H(F)$ of a sufficiently complex ANN can be used as least square estimators for posterior probabilities [17]. In our case the element $H_k(F)$ can be viewed as the probability of the input to belong to the k^{th} available class:

$$H : \mathbb{R}^M \rightarrow [0, 1]^N \in \mathbb{R}^N \quad (10)$$

We used M input nodes, a hidden layer of $(M + N)/2$ nodes and N output nodes. All activation functions are log-sigmoidal. The network was trained, using the scaled conjugate gradient method, a robust variant of the common back-propagation algorithm. In order to avoid over-fitting, the training session was early-stopped according to the increase of the classification error on 20% randomly chosen elements extracted from the training set.

C. Monte Carlo estimation

Given a density profile $F_{corrupted}$ with a corrupted part m, this latter part can be replaced by the corresponding data extracted from a synthetic profile of the k^{th} class, $g(i, \boldsymbol{s}_k^*)$, obtained as described in Sec. III-A, resulting in a new density profile F_k^*.

The probability $p(c_k|o)$, with $o = F_k^*$ that the profile belongs to the k^{th} class is evaluated for each class through the classification function $P(c|o) = [p(c_1|o), \ldots, p(c_K|o)] = H(F_k^*)$. By the imputation-classification procedure $I = 100$ times for each class, $K * I = 2400$ synthetic samples and classification vectors $P_i(c|o) = H(F_{k,i}^*)$ are generated.

The final classification is obtained by the Monte Carlo maximum a posteriori estimate:

$$\max_c E_i\left[P_i(c|o)\right] = \max_c \frac{1}{I}\sum_{i=1}^{I} P_i(c|o) \quad (11)$$

D. Class Means Imputation

At variance with the generative model approach described, a simpler imputation strategy is considering the mean profile for each class $\bar{F}$ obtained from a training set, and then obtaining the new profile F_k^* by replacing the corrupted part of $F_{corrupted}$ with the corresponding part in $\bar{F}$. Since no variability is taken into account, only K imputation can be performed, one for each different class, thus obtaining $P_k(c_k|o) = H(F_k^*)$. The final classification becomes:

$$\max_c E_i \left[P_k(c_k|o)\right] = \max_c \frac{1}{K} \sum_{k=1}^{I} P_k(c_k|o) \qquad (12)$$

IV. Results

The classification results are shown in Tab. I. For sake of comparison, the performance of the NN on the training set, comprising both corrupted and uncorrupted chromosomes, are also reported, along with the performance on the test data for the uncorrupted chromosomes alone. The results of the classification on the corrupted profiles in the test set, and on the instances using the class means imputation and on those obtained with the Monte Carlo strategy, are then shown. It can be clearly appreciated the improvement of the class mean imputation with respect to using the original corrupted features, and the further increase in performance by using the proposed strategy. It is worth noting that even if all probabilities of correct classification are below 0.5, they are far better than random guess for the 24-class problem considered, whose difficulty can be appreciated by comparing the original performance of the neural-network on the uncorrupted profiles (probability of correct classification of 0.87), with that on the corrupted profiles (probability of correct classification of 0.17).

TABLE I
Performance on Real Cytogenetic Data.

Dataset	Mean (StD)
Train - all chromosomes	0.87 (0.01)
Test - uncorrupted	0.84 (0.01)
Test - corrupted	0.17 (0.02)
Mean sub. - corrupted	0.28 (0.02)
MC sub. - corrupted	0.36 (0.01)

V. Conclusions

In this paper, we have shown that if a generative model for the data is available or if can be estimated from the data, it can be exploited in order to classify objects with corrupted or missing data, that are usually affected by a high classification error. By replacing the corrupted data with synthetic profiles obtained using the available generative models for each class, and under the assumption of uniform a priori probability, we transform the classification with missing data in a voting strategy among the profiles with the replaced data.

We show the superiority of the proposed approach in classifying real cytogenetic corrupted data both with respect to the imputation of the class-means profiles and to the classifier trained on the corrupted and uncorrupted data together, more than doubling the original classification performance.

Acknowledgment

The authors wish to thank TesiImaging S.r.l. for having kindly provided chromosome images and manual karyotypes.

Conflict of Interest

All authors have no conflict of interest.

References

[1] Internationa Standing Committee on Human Cytogenetic Nomenclature, *ISCN: an international system for human cytogenetic nomenclature (2005)*, L. G. Shaffer and N. Tommerup, Eds. Karger and Cytogenetics and Genome Research, 2005.

[2] J. Piper, "Genetic algorithm for applying constraints in chromosome classification," *Pattern Recognition Letters*, vol. 16, pp. 857–864, 1995.

[3] M. Moradi and S. K. Staredhan, "New features for automatic classification of human chromosomes: a feasibility study," *Pattern Recognition Letters*, vol. 27, pp. 19–28, 2006.

[4] B. Lerner, H. Guterman, I. Dinstein, and Y. Romem, "Medial axis transform-based features and neural network for human chromosome classification," *Pattern Recognition*, vol. 28, pp. 1673–1683, 1995.

[5] E. Grisan, E. Poletti, and A. Ruggeri, "Automatic segmentation and disentangling of chromosomes in q-band prometaphase images." *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, pp. 575–581, April 2009.

[6] G. Agam and I. Dinstein, "Geometric separation of partially overlapping nonrigid objects applied to automatic chromosome segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 1212–1222, November 1997.

[7] T. Hastie, R. Tibishirani, and J. Friedman, *The elements of statistical learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer, 2001.

[8] R. Little and D. Rubin, *Statistical Analysis with Missing Data*. New York, NY, USA: Wiley, 1987.

[9] M. Tanaka, Y. Kotokawa, and T. Tanino, "Pattern classification by stochastic neural network with missing data," in *Systems, Man, and Cybernetics, 1996., IEEE International Conference on*, vol. 1, Oct. 1996, pp. 690 –695 vol.1.

[10] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *In Advances in Neural Information Processing Systems 11*. MIT Press, 1998, pp. 487–493.

[11] E. Poletti, E. Grisan, and A. Ruggeri, "Automatic classification of chromosomes in q-band images," in *Proc. 30th Annual International IEEE EMBS Conference*, Vancouer, BC, Canada, August 2008.

[12] A. McQuarrie, "The model selection criterion aicu," *Statistics & Probability Letters*, vol. 34, no. 3, pp. 285–292, 1997.

[13] B. Lerner, H. Guterman, and I. Dinstein, "A classification-driven partially occluded object segmentation (CPOOS) method with application to chromosome analysis," *IEEE Transactions on Signal Processing*, vol. 46, no. 10, pp. 2841–2847, October 1998.

[14] N. Sweeney, R. Becker, and B. Sweeney, "A comparison of wavelet and fourier descriptors for a neural network chromosome classifier," in *Proceedings of the 19th International Conference IEEE/EMBS*, Chicago, November 1997, pp. 1359–1362.

[15] S. Delshdapour, "Reduced size multi layer perceptron neural network for human chromosome classification," in *The 25th Silver Anniversary International Conference of the IEEE Engineering in Medicine and Biology Society*, Cancun (Mx), September 2003, pp. 2249–2252.

[16] A. M. Badawi, K. G. Hasan, E. A. Aly, and R. A. Messiha, "Chromosomes classification based on neural networks, fuzzy rule based, and template matching classifiers," in *Proceedings of The 46th IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2003.

[17] M. Hung, M. Hu, M. Shanker, and P. Patuwo, "Estimating posterior probabilities in classification problems with neural networks," *International Journal fo Computational Intelligence and Organizations*, vol. 1, pp. 49–60, 1996.