ELSEVIER

Artificial Intelligence in Medicine 30 (2004) 301–316

Artificial Intelligence in Medicine

www.intl.elsevierhealth.com/journals/aiim

Bayesian fluorescence in situ hybridisation signal classification

Boaz Lerner*

Pattern Analysis & Machine Learning Lab, Department of Electrical & Computer Engineering, Ben-Gurion University, Beer-Sheva, Israel

Received 2 June 2002; received in revised form 8 August 2003; accepted 10 August 2003

Abstract

Previous research has indicated the significance of accurate classification of fluorescence in situ hybridisation (FISH) signals for the detection of genetic abnormalities. Based on well-discriminating features and a trainable neural network (NN) classifier, a previous system enabled highly-accurate classification of valid signals and artefacts of two fluorophores. However, since this system employed several features that are considered independent, the naive Bayesian classifier (NBC) is suggested here as an alternative to the NN. The NBC independence assumption permits the decomposition of the high-dimensional likelihood of the model for the data into a product of one-dimensional probability densities. The naive independence assumption together with the Bayesian methodology allow the NBC to predict a posteriori probabilities of class membership using estimated class-conditional densities in a close and simple form. Since the probability densities are the only parameters of the NBC, the misclassification rate of the model is determined exclusively by the quality of density estimation. Densities are evaluated by three methods: single Gaussian estimation (SGE; parametric method), Gaussian mixture model assuming spherical covariance matrices (GMM; semi-parametric method) and kernel density estimation (KDE; non-parametric method). For low-dimensional densities, the GMM generally outperforms the KDE that tends to overfit the training set at the cost of reduced generalisation capability. But, it is the GMM that loses some accuracy when modelling higher-dimensional densities due to the violation of the assumption of spherical covariance matrices when dependent features are added to the set. Compared with these two methods, the SGE and NN provide inferior and superior performance, respectively. However, the NBC avoids the intensive training and optimisation required for the NN, demanding extensive resources and experimentation. Therefore, when supporting these two classifiers, the system enables a trade-off between the NN performance and NBC simplicity of implementation.

Keywords: Bayesian network; Density estimation; Fluorescence in situ hybridisation (FISH); Gaussian mixture model; Naive Bayesian classifier; Signal classification

* Tel.: +972-8-6472567; fax: +972-8-6472949.
E-mail address: boaz@ee.bgu.ac.il (B. Lerner).

0933-3657/$ – see front matter
doi:10.1016/j.artmed.2003.11.005

1. Introduction

Fluorescence in situ hybridisation (FISH) allows selective staining of various DNA sequences in interphase nuclei and thereby the detection, analysis and quantification of specific numerical and structural chromosomal abnormalities within these nuclei.

Digital microscopy in FISH enables the application of image analysis techniques for automation of time consuming tasks, such as dot counting. Dot (signal) counting is considered one of the most important applications of FISH as the signals represent the inspected DNA sequences. The common approach to dot counting relies on an auto-focusing mechanism to select the most focused image for the analysis [15]. However, basing dot counting on auto-focusing has some shortcomings [12]. Instead, it has recently been proposed [12] to base FISH dot counting on a classifier that discriminates between in and out of focus images captured at different focal planes of the same field of view (FOV). Each image is analysed and its signals are discriminated by the classifier as valid data or artefacts, which are the result of out-of-focusing. The image that contains no artefacts is selected as the in-focus image to represent that FOV. The method overcomes most of the shortcomings of auto-focusing since it utilises three rather than two-dimensional cell information. However, since the system captures images that contain many more unfocused signals, its accuracy in distinguishing focused and unfocused signals should be superior to that of a system employing an auto-focusing mechanism. Therefore, the developed system is based on the extraction of well-discriminating characteristics of focused and unfocused signals [13], and accurate classification of these signals into real (valid) signals and artefacts, respectively.

In previous work [12], we suggested a generic FISH signal classification methodology. Several hierarchical neural network (NN) strategies were investigated leading to low classification errors. Since on the one hand, NN training and architecture optimisation need a large amount of resources and experimentation, and on the other hand the suggested methodology is independent of classifier type, we extend our study here to evaluate the Bayesian network (BN) classifier in FISH signal discrimination. A BN encompasses a graphic representation of dependencies between problem variables (the structure) together with Bayesian evaluation of probabilities quantifying these dependencies. The BN impressive interpretability and adaptivity in learning structures compared with other machine learning paradigms (e.g. NNs) have established its importance in tackling classification problems [4,6]. The naive Bayesian classifier (NBC) [7,11] is a type of BN restricted to variables independent of each other given the class variable. As a simple alternative to the NN, the NBC employs the variable independence assumption along with the Bayesian formalism to update prior to posterior probabilities of class membership using the likelihood function. Then, an unseen pattern is assigned to the class having the highest posterior probability. Previous research [13] has evaluated FISH signal features and indicated independence among some of them which is employed here as a priori information for the construction of the NBC. Unfortunately, most previous work utilising BNs has dealt only with discrete variables or when coping with continuous variables either discretized the variables or assumed they can be explained by a single Gaussian data generator [8,11,18,19]. Since most of the FISH signal features are continuous, we compute and compare the likelihood for these features estimated using parametric, non-parametric and semi-parametric methods.

Section 2 describes the procedure we employ to acquire FISH images, while Section 3 depicts a methodology for multi-spectral FISH image analysis and signal measurement. Section 4 presents the naive Bayesian classifier applied to FISH signal classification. Section 5 exemplifies a parametric, semi-parametric and non-parametric estimation methods for the likelihood of the model, namely single Gaussian estimation (SGE), Gaussian mixture model (GMM) and kernel density estimation (KDE), respectively. Finally, Section 6 demonstrates the experimental study and its results, while Section 7 summarises the work.

2. FISH image acquisition

Interphase nuclei preparations from amniotic fluid were made using the method by Klinger et al. [10] with minor modifications. Cells were put directly onto slides and target areas were marked. Slides were screened under a Zeiss axioplan epifluorescence microscope using Zeiss 100× objective. Signals were viewed using Vysis DAPI/green/orange triple bandpass filter set and images acquired using a CCD camera (Photometrics CH250/A) and SmartCapture software (Vysis, Downers Grove, IL). Red and green signals, corresponding to chromosomes 21 and 13, respectively, were seen on blue DAPI stained nuclei. Since manual acquisition of stacks of images in different focal planes for the different FOVs is a relatively demanding task, we follow in this work a simpler procedure. Slides were scanned by starting in the upper left corner of the coverslip and moving from top to bottom. The focus and colour ratios were adjusted for the first captured image from each slide, and then kept at those values for all the following images from that particular slide. Images were captured by stopping at random intervals along the slide. Utilising this acquisition procedure and assuming uniform distribution of signals along the Z-axis, we captured an arbitrary mix of in-focus and out-of-focus images without literally collecting stacks of images. For the purpose of evaluating the classification of focused and unfocused signals, this procedure provides the desired example images cheaply and quickly, but for testing the entire system in dot counting stacks of images should be acquired. A total of 400 in-focus and out-of-focus images were collected from five slides, stored in TIFF (Tagged Image File Format) format and used in the signal classification experiments.

3. Colour image analysis and signal measurement

In FISH preparation, multiple probes labelled by different fluorophores are frequently combined. In the current study for instance, chromosomes 13 and 21 are detected as green and red signals, respectively, whereas the nuclei are indicated by blue. By analysing each of the three colour channels—red, green and blue (RGB)—of a FISH image separately, image processing can be facilitated. Nuclei are analysed using the blue channel of the RGB image, whereas the signals are analysed using the red and green channels. Segmentation on each of the three channels using global thresholds yields the image nuclei and red and green signals. Noise elimination and boundary smoothing of nuclei, as well as spatial correlation between nuclei and signals, complete the segmentation [13].

Multi-spectral FISH image analysis is beneficial not only to facilitate pre-processing and segmentation, but also to yield colour-based features that contribute to an efficient signal classification [13]. In this work, RGB colour format is utilised following image acquisition because pre-processing, as well as nuclei and signal segmentation, are performed more easily using this colour format than using the conventional conversion of the image to grey-level scale. However, as intensities of red and green signals, each measured in its own channel, are very similar to each other, the RGB format is not suitable for discriminating between signals of different colours. By contrast, signals of different fluorophores represented by the hue parameter of the HSI (hue, saturation, intensity) colour format [17] can be easily resolved due to their different hue. Multi-spectral image analysis may also alleviate classification of signals of different fluorophores being in close proximity to each other as this leads to ambiguity and thus accuracy deterioration.

Following segmentation, features are measured for the signals. Features include area (a size measure) and eccentricity (a shape measure), which have been previously suggested [15]. In addition, we measure a number of spectral features. We compute at the specific colour plane three RGB intensity-based measurements: the total and average channel intensities and the channel intensity standard deviation. We also compute four HSI hue-based measurements: maximum hue, average hue, hue standard deviation, and delta hue. Delta hue is the difference between the maximum and average hue normalised by the average hue. This feature has been added to the set because it was observed that the difference between values of the average and maximum hue for real signals is usually near zero, whereas artefacts having mixed colour components get substantially large values for this difference. Two additional features of the set are the two co-ordinates of the eigenvector corresponding to the largest eigenvalue of the red and green intensity components of the signal. The last feature is the average grey intensity, i.e. average intensity over the three colour channels (motivation for choosing the last four features is given in [13]). Table 1 lists and numbers the 12 features to facilitate their identification in the rest of the paper.

Table 1
The set of features studied in the work

Number	Feature
1	Area
2	Eccentricity
3	Total channel intensity
4	Average channel intensity
5	Texture
6	Maximum hue
7	Average hue
8	Hue texture
9	Delta hue
10	Eig. 1
11	Eig. 2
12	Average grey intensity

Numbers are used in the rest of the paper to identify the features. Texture indicates standard deviation of intensity (5) or hue (8). Eig. 1 and 2 are abbreviations for the two co-ordinates of the eigenvector corresponding to the largest eigenvalue of the red and green intensity components of the signal.

4. The naive Bayesian classifier

A BN for a finite set $\boldsymbol{U} = \{X_1, X_2, \ldots, X_n\} = \{\boldsymbol{X}\}$ of random variables consists of a network structure $\boldsymbol{S}$ that encodes a set of conditional independence assertions about variables in $\boldsymbol{X}$, and a set $\boldsymbol{P}$ of local probability distributions associated with each variable. Together, the two define the joint probability over $\boldsymbol{X}$. The network structure $\boldsymbol{S}$ is a directed acyclic graph in which the nodes represent the variables $\boldsymbol{X}$. We use X_i to denote both the variable and its corresponding node, and Pa_i to denote parent nodes of node X_i as well as the variables corresponding to these parents. Arcs and missing arcs encode dependencies and independencies, respectively, in $\boldsymbol{S}$. Given structure $\boldsymbol{S}$, the joint probability distribution for $\boldsymbol{X}$ is given by [6]

$$p(\boldsymbol{X} = \boldsymbol{x}) = \prod_{i=1}^{n} p(X_i = x_i | \mathrm{Pa}_i) \tag{1}$$

where $\boldsymbol{x}$ is the assignment of states to each variable in $\boldsymbol{X}$, x_i the value taken by X_i, and the terms in the product compose the set of local probability distributions $\boldsymbol{P}$.

The NBC is a special case of a BN consisting of a finite set of random variables, $\boldsymbol{U} = \{X_1, X_2, \ldots, X_m, C\} = \{\boldsymbol{X}, C\}$, where $X_1, \ldots, X_m$ are the observable variables that represent the features, and C the class variable having K states. The classifier provides a simple method of pattern classification, while still enabling impressive performance. The NBC is termed naive since it makes use of a simplifying assumption that its observable variables are conditionally independent given the class variable. All arcs of the NBC are directed from the class variable to the observable variables (Fig. 1), hence $\mathrm{Pa}_i = C$ for each of the observable variables and $\mathrm{Pa}(C) = \emptyset$ for the class variable.

The NBC assigns a test pattern $\boldsymbol{x}$ to the class C_k $(k = 1, \ldots, K)$ with the highest a posteriori probability

$$P(C_k | \boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)\, P(C_k)}{p(\boldsymbol{x})} \tag{2}$$

where $p(\boldsymbol{x}|C_k)$ is the class-conditional probability density, $P(C_k)$ the a priori probability for class C_k, and $p(\boldsymbol{x})$ the unconditional density normalising the product of the former two such that $\sum_k P(C_k|\boldsymbol{x}) = 1$. Using the NBC independence assumption and omitting

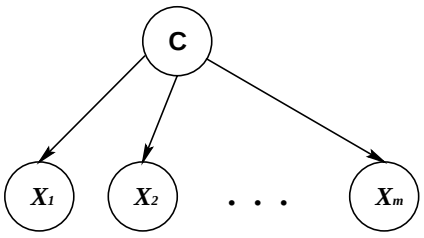

Fig. 1. The naive Bayesian classifier depicted as a Bayesian network in which the observable variables $(X_1, X_2, \ldots X_m)$ are conditionally independent given the class variable (C).

$p(\boldsymbol{x})$ which is common to all states of the class variable, the posterior probability can be written as

$$P(C_k|\boldsymbol{x}) \propto p(\boldsymbol{X} = \boldsymbol{x}|C_k)\, P(C_k) = P(C_k) \prod_{i=1}^{m} p(X_i = x_i|C_k) \tag{3}$$

where $\boldsymbol{X} = \boldsymbol{x}$ represents the event that $X_1 = x_1 \wedge X_2 = x_2 \wedge \cdots \wedge X_m = x_m$ and $\prod_{i=1}^{m} p(X_i = x_i|C_k)$ is a product of class-conditional densities for $\boldsymbol{x}$. Both $P(C_k)$ and $p(X_i|C_k)$ can be estimated from the training data; the estimation of $p(X_i|C_k)$ is described in Section 5, while $P(C_k)$ is the relative frequency of patterns belonging to C_k out of all of the patterns in the data. Classification of focused (real) and unfocused (artefact) red and green FISH signals represented by the features of Table 1 is performed by the NBC class variable taking up four values: 'real red', 'artefact red', 'real green' and 'artefact green'.

5. Estimation of class-conditional densities

Introducing the independence assumption, conditional densities of the NBC decompose (Eq. (3)). This eliminates the ''curse of dimensionality'' since density estimation requires only linearly rather than exponentially increasing numbers of patterns. We need to estimate $p(X_i|C_k)$, the one-dimensional class-conditional probabilities (for discrete variables) or probability densities (for continuous variables) for each class C_k and variable X_i. For this purpose, we employ a training set comprising of a finite number of data points $\boldsymbol{x}^n$, where n gets values for each of the N_k training patterns of class C_k.

Given a selected network structure, the parameters determining the accuracy of the NBC are the class-conditional probabilities/densities. For a discrete variable, the class-conditional probability is estimated using the sample frequency of each value of the variable. For a continuous variable, an evaluation of density estimation methods, assuming different data generation mechanisms, is essential in order to decide on the most accurate method. Since class-conditional probability densities are usually modelled by parameterised functional forms, $p(x|C_k)$ are referred to as *likelihood* functions for x and the parameters are estimated using the maximum likelihood procedure. In this study, we model the probability distribution of the area feature, which is the only discrete feature in Table 1, using its sample frequency, and estimate the probability densities for all continuous features using three methods assuming different mechanisms of data generation. Single Gaussian estimation assumes the data are generated from a single normal distribution, whereas kernel density estimation models the data using a linear combination of kernels around each of the training samples. The Gaussian mixture model estimates the data using a few Gaussians with adaptable parameters.

5.1. Single Gaussian estimation

In most previous work that has dealt with continuous variables of an NBC, data were either discretized [11] or assumed to be generated by a parametric model that is based on a

single Gaussian distribution [7]. The assumption of normal distribution has convenient analytical and statistical properties and it is suitable in representing measurements of many natural phenomena. For each of the one-dimensional class-conditional densities of the NBC, the normal density function for x can be written as

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \tag{4}$$

where μ and σ are the mean and standard deviation of the distribution, respectively. The SGE of the mean and standard deviation of the normal distribution measured by the maximum likelihood procedure are

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x^n \tag{5}$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (x^n - \hat{\mu})^2 \tag{6}$$

where N (N_k in the case of class C_k) is the number of training patterns x^n. This is the intuitive result that the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}$ of the mean and standard deviation μ and σ of the distribution are given, respectively, by the sample average and standard deviation.

5.2. Kernel density estimation

Non-parametric techniques for probability density estimation do not specify the functional form of the density beforehand but use the data to estimate it. One of the most common approaches to non-parametric modelling is kernel density estimation. KDE models the one-dimensional density as

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h} H\left(\frac{x - x^n}{h} \right) \tag{7}$$

using kernel functions

$$H(t) = \begin{cases} 1, & |t| < \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

having width parameter h and centred around each of the training data points x^n. Estimating the width parameter is usually performed by maximising the likelihood function [1,20] or minimising least squares or mean square error [5,16,20]. Alternatively, h may be modelled using a parametric form such as $h = TN^{\alpha}$ [5,7,20], where $T > 0$, $-1 < \alpha < 0$ (typically $\alpha = -1/2$ [3]) and $N = N_k$ for each class C_k. This choice guarantees that the parameter shrinks to zero as the number of points goes to infinity, and hence KDE becomes increasingly local as the number of training points increases.

By introducing normal kernel functions, we can overcome discontinuities in the model, so

$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{1/2}}\exp\left\{-\frac{\| x - x^n \|^2}{2h^2}\right\} \tag{9}$$

for each class-conditional density. A typical value of α for normal kernel functions is $-1/5$ [20]. Other (smoothing) kernels, satisfying the conditions of a probability density function may also be employed [3,20].

Non-parametric models can replace parametric models, which may not hold for domains where data are not normally distributed. For this reason, it is suggested in [7] to replace SGE with KDE when modelling the class-conditional densities of the NBC. Non-parametric methods model non-normal distributed data more accurately than parametric techniques but at the cost of storage and computational complexities as the number of variables in the model grows linearly with the number of training data points.

5.3. Gaussian mixture model

Semi-parametric methods try to combine the benefits of both parametric and non-parametric methods. They are not restricted to specific functional forms, and yet the model size depends only on the problem complexity and not on the data size.

A GMM is one of the most powerful semi-parametric techniques. It estimates the data density using a linear combination of basis functions similarly to KDE. However, the number of basis functions M is a parameter of the model, which is much less than the number N of data points. Based on a linear combination of one-dimensional component densities $p(x|j)$ with some mixing coefficients $P(j)$, also called the prior probabilities, the model for the density is [1]

$$p(x) = \sum_{j=1}^{M} p(x|j)\, P(j) \tag{10}$$

where $P(j)$ satisfy the probability constraints, i.e. $\sum_{j=1}^{M} P(j) = 1$ and $0 \leq P(j) \leq 1$, and $p(x|j)$ are normalised so that $\int p(x|j)\,\mathrm{d}x = 1$.

Assuming the component densities are Gaussian distribution functions with means μ_j and standard deviations σ_j, the density is modelled by

$$p(x) = \sum_{j=1}^{M}\frac{1}{(2\pi\sigma_j^2)^{1/2}}\exp\left\{-\frac{\| x - \mu_j \|^2}{2\sigma_j^2}\right\}P(j). \tag{11}$$

Now, besides estimating μ_j and σ_j we should also estimate $P(j)$ simultaneously. Most of the methods for determining these parameters from the data are based on the maximum likelihood procedure. One such method, which makes use of the expectation-maximisation (EM) algorithm [2], is employed in the experiment described in Section 6. More details on the EM algorithm can be found in [1,2].

6. Experimental study

Generally, to induce a Bayesian network for classification, we first have to find an optimal network structure. We can either use prior knowledge or employ the data [4,6] to select a structure or average over a several structures in order to find the optimal network. For FISH signal classification, such prior knowledge suggests [13] that several feature representations contribute the most for accurate classification of the signals into the four classes of Section 4. Thus, structure nodes of our BN model should represent variables corresponding to these features along with the class variable. In the case of an NBC, finding a structure is trivial as variables are independent of each other given the class variable. Then, we only need to estimate the class-conditional probability densities (or distributions) for each variable given each of the four states of the class variable.

We estimate the class-conditional probability densities using SGE, KDE and GMM. For the KDE, we employ $\alpha = -1/2$ and cross-validation (CV) to estimate on a validation set the optimal T for the calculation of h (for details see [14]). For the GMM, 10 Gaussians are selected based on preliminary visualisation of the data in order to estimate densities of each of the features. In order to reduce the computational load, spherical covariance matrices for all Gaussians are assumed. Convergence of the EM algorithm is identified whenever the difference in error between 2 consecutive EM iterations is smaller than a predefined threshold, and training, usually following 10 iterations, is stopped.

Figs. 2 and 3 show, respectively, two examples of class-conditional probability densities for the average channel intensity and average hue features given the four states of the class variable when estimated by the three methods. Figs. 2 and 3 demonstrate that data modelling using KDE is frequently spikier (especially for the artefact classes) than using SGE and GMM, since it depends on the actual training data points heavily rather than the general mechanism that has generated the points. Moreover, this tendency of the KDE leads the model to overfitting the training set. The KDE and GMM peaks follow mass centres of distributions accurately and the SGE variance is large in order to capture the entire distribution. In addition, differences among estimation methods are more evident for the 'artefact' classes in which feature values are more distributed than for the 'real' classes.

The experiments to evaluate the NBC accuracy are conducted using 10-fold cross-validation (CV-10). Tables 2 and 3 show classification accuracy on the test set for single features and sub-sets of features, respectively, when the class-conditional densities are modelled using SGE, KDE and the GMM. Each sub-set of features is responsible for a different configuration of the NBC. Tables 2 and 3 reveal that estimating density for the NBC by the GMM is advantageous to the KDE for low-dimensional densities ($m = 1, 2$) and vice versa for high-dimensional densities. The inferiority of the KDE for low-dimensional densities is attributed to the model inherent overfitting of the training set deteriorating the generalisation capability for the FISH data. On the other hand, the inferiority of the GMM for high-dimensional densities of growing numbers of dependent features is attributed to increasing violation of the implementation assumption of spherical covariance matrices.

Both models outperform the SGE in all cases sometimes even dramatically. For example, the accuracy of the NBC when density is modelled by the SGE for the delta

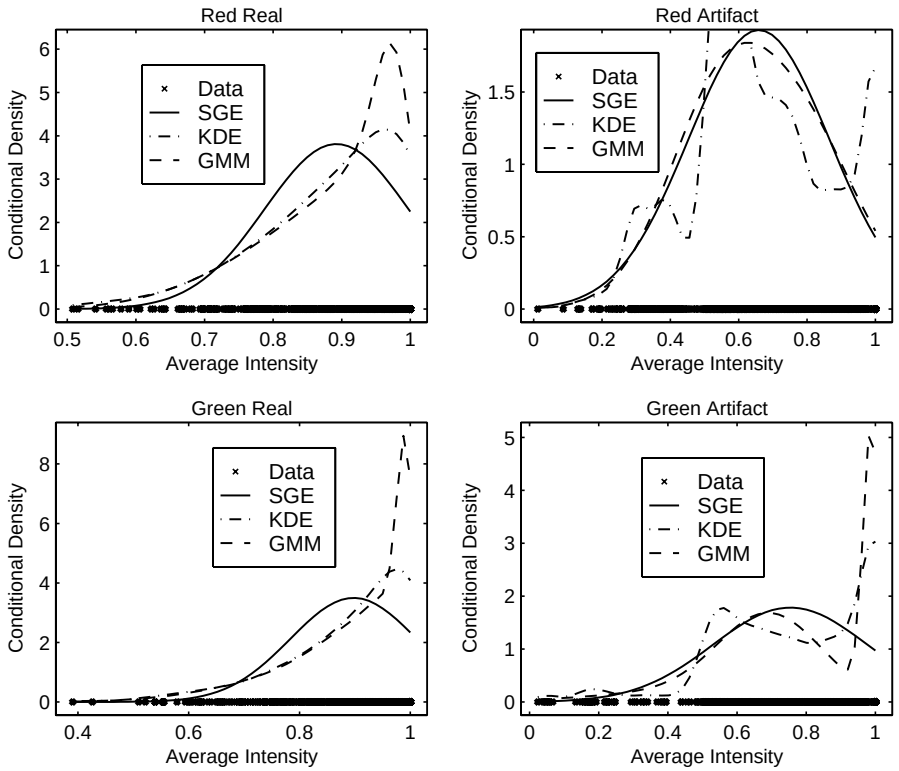

Fig. 2. Class-conditional density estimation of the average channel intensity feature given each of the four states of the class variable. Modelling is performed using three methods—single Gaussian estimation, kernel density estimation and the Gaussian mixture model.

hue feature (9) is around 40% less than that for the KDE and GMM (Table 2). This inferiority of the SGE may be understood by comparing the true and estimated probability density functions (pdfs) for this feature in Figs. 4 and 5, respectively. The comparison explains the poor modelling of the SGE by its overlapped representation when based on

Table 2
The naive Bayesian classifier accuracy (%), represented by the mean and standard deviation when the class-conditional probability densities for single features are estimated by SGE, KDE and GMM

Feature number	SGE	KDE	GMM	MLP
7	60.5 (2.8)	63.4 (1.7)	69.2 (2.6)	69.5 (1.9)
6	62.7 (2.1)	56.0 (2.7)	65.4 (2.2)	64.4 (2.8)
12	47.4 (3.2)	47.2 (3.3)	47.7 (2.9)	47.2 (3.1)
4	44.8 (2.6)	45.4 (2.6)	44.3 (2.7)	45.2 (2.1)
9	27.9 (2.5)	45.4 (1.9)	45.1 (2.8)	47.5 (1.6)
1	45.2 (2.8)	45.1 (2.8)	45.1 (2.8)	45.1 (2.8)

Feature numbers are defined in Table 1. The probability distribution is computed for the area feature (1). The accuracy is compared to that achieved by a multilayer perceptron (MLP) neural network [13].

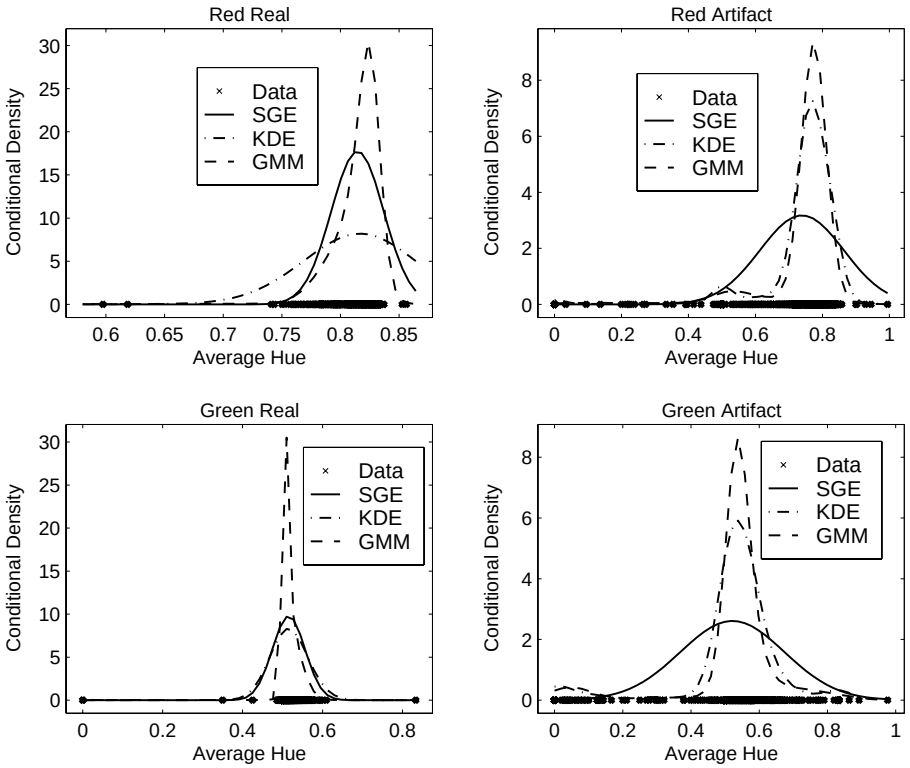

Fig. 3. Class-conditional density estimation of the average hue feature given each of the four states of the class variable. Modelling is performed using the three methods as in Fig. 2.

delta hue. Although its improved accuracy the KDE suffers from a known computational flaw. For m features, N training samples and M test samples, the KDE must perform $O(N^2m)$ and $O(nNM)$ evaluations during the training and the test, respectively, where SGE and GMM need only $O(mN)$ and $O(mM)$ evaluations, respectively.

Table 3
The naive Bayesian classifier accuracy similar to Table 2, but for multivariate feature representation

Feature combination	SGE	KDE	GMM	MLP
4, 12	48.4 (3.3)	48.7 (3.2)	48.6 (3.1)	68.6 (3.1)
1, 7	75.7 (2.1)	77.9 (2.8)	78.9 (1.8)	78.0 (2.2)
1, 3, 4	39.4 (4.7)	48.1 (2.8)	49.9 (4.5)	80.1 (2.1)
1, 7, 12	71.8 (2.1)	79.1 (2.5)	77.1 (1.5)	82.0 (2.0)
1, 4, 6	73.7 (2.3)	77.5 (2.3)	75.7 (2.9)	80.8 (2.2)
4, 8, 12	46.1 (3.4)	48.8 (3.2)	44.8 (4.0)	82.0 (2.5)
3, 4, 8, 12	44.0 (3.9)	48.3 (2.8)	45.2 (4.3)	82.1 (2.5)
1, 6, 9, 12	65.0 (1.4)	79.2 (2.2)	71.1 (4.6)	82.3 (2.5)
All	68.8 (3.1)	78.7 (2.5)	66.5 (5.8)	84.9 (1.7)

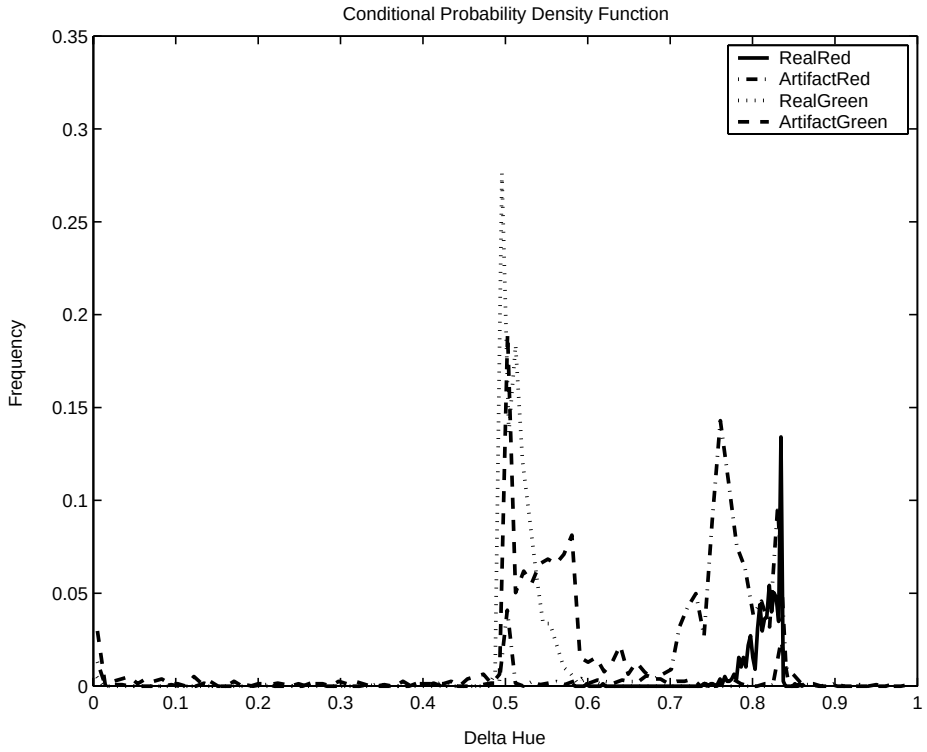

Fig. 4. The true class-conditional probability density function of the delta hue (9) feature for each of the four states of the class variable.

The accuracy of the classifier can also be used to evaluate network configuration and thereby select the most discriminative features. It is demonstrated in Tables 2 and 3 that some manually selected features drawn from Table 1, such as the average (7) and maximum hue (6), as well as the area (1) and average grey intensity (12), enable well-discriminative representations of the signals (a rigorous feature selection method is suggested in [13]). Furthermore, Table 3 compares sub-sets of features differ in numbers (2–12) and content. It is revealed that the most dominant factor determining the accuracy of the NBC, regardless of which estimation method is selected, is the amount of dependency contained within the feature sub-set. High degree of feature dependence leads to accuracy deterioration as the independence assumption being violated. For example, correlation coefficients $\rho = 0.83$ and 0.99 are measured between features 4 (average channel intensity) and 12 (average grey intensity) and features 1 (area) and 3 (total channel intensity), respectively. These high coefficients are reflected in the NBC low accuracy when estimation is based on sub-sets {4, 12}, {4, 8, 12} and {1, 3, 4}. In contrast, pairs of features having low correlation coefficients, such as 1 and 7 (average hue) ($\rho = 0.071$) and 1 and 12 ($\rho = 0.0034$), are associated in sub-sets {1, 7} and {1, 7, 12} with relatively high classification accuracy.

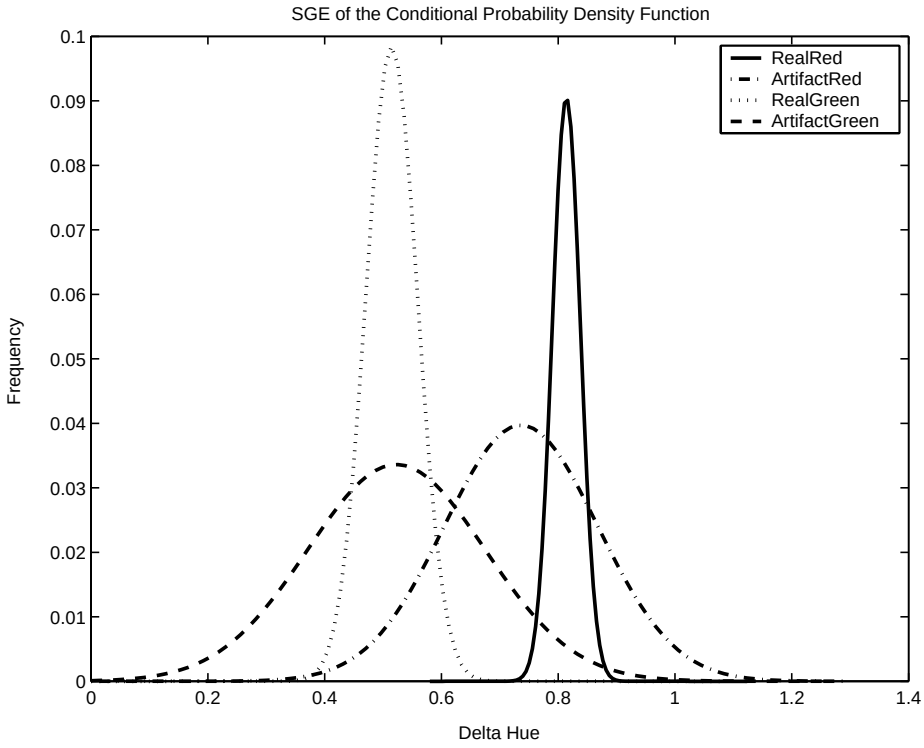

Fig. 5. The single Gaussian estimated class-conditional probability density function of the delta hue feature for each of the four states of the class variable.

Finally, the NBC accuracy is compared to that of the multilayer perceptron (MLP) neural network [1], which was evaluated for this task previously [12,13]. Represented by single features, signals are discriminated similarly by the two classifiers (Table 2). Thus, any difference in accuracy between the two classifiers discriminating the same signal high-dimensional feature representation (Table 3) should be attributed to the characteristics of the classifiers and not to the features themselves. Table 3 shows that the MLP outperforms the NBC regardless of the feature sub-set or density estimation method used by the NBC. As the number of dependent features grows in the set, this superiority increases (see e.g., feature sub-sets {3, 4, 8, 12} and 'All'). This is due to the increased violation of the independence assumption by the NBC, whereas the MLP hidden layer extracts better data representations for larger feature sets even if they contain some correlated features.

7. Discussion

Accurate FISH signal classification was found essential for dot counting, a common procedure necessary in detecting genetic abnormalities such as Down Syndrome.

Also found that signal classification in FISH image analysis is inevitable when the images are captured without relying on an auto-focusing mechanism or acting as an alternative to auto-focusing. Thus, in the current study, we evaluate a special Bayesian network, namely the naive Bayesian classifier, in classifying FISH signals.

A BN classifier consists of two elements. The first is network structure, which is determined in this study for the NBC using a priori knowledge of the possible states of the class variable and the significance of features representing FISH signals. That is, the most significant features are represented by the observable variables which are assumed to be independent given the class variable. The second component of the classifier is a set of parameters that quantify the structure. Generally, efforts are concentrated in improving network structure [4,6,8,11,18,19] neglecting parameter learning. In this study, however, we have focused on modelling the NBC parameters, i.e. estimating the class-conditional probability densities of each observable variable given each state of the class variable. Then, the densities are employed using the Bayesian formalism to assign an unseen signal pattern to the class having the highest a posteriori probability.

Three approaches of density estimation are considered in this work. The first approach uses a parametric method in which a specific functional form, namely normal distribution, is assumed for modelling the density. Maximum likelihood procedure provides the estimated mean and standard deviation for this single Gaussian estimation as the sample average and standard deviation, respectively. Non-parametric estimation techniques, like kernel density estimation, do not assume a particular functional form, but allow the form of the density to be determined entirely by the data. However, as the number of parameters in the model and the number of evaluations grow (linearly and quadratically, respectively) with the data size, the non-parametric model can quickly become unwieldy. Furthermore, the problem of outliers which worsens the accuracy of non-parametric methods increases with dimensionality. Semi-parametric estimation methods, such as Gaussian mixture model, try to achieve the best of both approaches. They permit a very general class of functional forms in which more flexible models can be built by increasing the number of adaptive parameters. This number of parameters, however, can be varied independently from the data size.

The three approaches are evaluated in density estimation by the NBC accuracy. As the accuracy of the classifier having a specific structure is determined by the quality of estimation of the class-conditional densities, an emphasis has been made to model and compare data generation assuming different mechanisms. Moreover, employing classification accuracy to evaluate density estimation methods utilised for classification is advantageous to other criteria since it is the same criterion maximised during the classification.

The accuracy of the NBC when densities being estimated by KDE has been found inferior or superior to that based on the GMM for low- or high-dimensional densities, respectively. In the former case, the KDE lacks to generalise well following overfitting the training data, and in the latter case the GMM assumption of spherical covariance matrices breaks as more dependent features are included in the set. Accuracy based on either model outperforms that based on SGE but is inferior to that of the MLP neural network, which is another non-parametric estimation method. This inferiority is attributed to the (naive) assumption of independence of the NBC and to the MLP architecture. The MLP combines

in its output layer (class variables) representations derived from hidden units extracting different non-linear representations of the same features that enable richer and more accurate modelling of the feature space. However, this advantage is reduced when feature dependency is weakened.

Accurate modelling of high-dimensional probability densities is extremely difficult. By making use of the NBC conditional independence assumption we eliminate this problem. We decompose the density into a product of one-dimensional densities and estimate each density separately. Employing features that are known to be independent of each other can therefore exploit the simplicity and accuracy of the NBC. In other cases where features are found to be dependent, like in this study, the independence assumption is violated and the accuracy of the NBC decreases. For these cases, the assumption may be relaxed and search algorithms, such as hill climbing [4,9] extending the structure of the NBC to include arcs corresponding to dependencies, can be employed. Usually, this extension will improve classification accuracy [4,18,19] but at the cost of losing simplicity.

Finally, the classification accuracy of the NBC is exploited here to evaluate and select features, thereby also performing model selection. This way, we facilitate and expedite structure learning while also performing parameter learning.

Acknowledgements

This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University, Beer-Sheva, Israel. The author thanks Roy Malka for his assistance during the preparation of this manuscript and to Seema Dhanjal for FISH image acquisition and labelling.

References

[1] Bishop CM. Neural networks for pattern recognition. Oxford: Clarendon Press; 1995.
[2] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B: Methodological 1977;39:1–38.
[3] Duda RO, Hart PE, Stork DG. Pattern classification. New York: Wiley; 2001.
[4] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learn 1997;29:131–63.
[5] Hall P, Johnstone I. Empirical functions and efficient smoothing parameter selection. J R Stat Soc Ser B: Methodological 1992;54:475–530.
[6] Heckerman D. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, March 1995 [revised November 1996].
[7] John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. In: Besnard P, Hanks S, editors. Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann; 1995. p. 338–45.
[8] Kahn Jr. CE, Roberts LM, Shaffer KA, Haddawy P. Construction of a Bayesian network for mammographic diagnosis of breast cancer. Comput Biol Med 1997;27:19–29.
[9] Keogh E, Pazzani M. Learning augmented Bayesian classifiers: a comparison of distribution-based and classification-based approaches. In: Proceedings of the Seventh International Workshop on AI and Statistics [Uncertainty 99], Ft. Lauderdale, FL, 1999. p. 225–30.
[10] Klinger K, Landes G, Shook D, Harvey R, Lopez L, Locke P, et al. Rapid detection of chromosome aneuploidies in uncultured amniocytes by using fluorescence in situ hybridisation (FISH). Am J Hum Genet 1992;51:55–65.

[11] Kohavi R, Becker B, Sommerfield D. Improving simple Bayes. In: Carbonell JG, Widmer G, Siekmann J, editors. The Poster Papers of the Ninth European Conference on Machine Learning. New York: Springer-Verlag; 1997. http://www.citeseer.nj.nec.com/kohavi97improving.html.
[12] Lerner B, Clocksin WF, Dhanjal S, Hultén MA, Bishop CM. Automatic signal classification in fluorescence in situ hybridization images. Cytometry 2001;43:87–93.
[13] Lerner B, Clocksin WF, Dhanjal S, Hultén MA, Bishop CM. Feature representation and signal classification in fluorescence in situ hybridization image analysis. IEEE Trans Syst Man Cybern A 2001; 31:655–65.
[14] Lerner B, Lawrence ND. A comparison of state-of-the-art classification techniques with application to cytogenetics. Neural Comput Appl 2001;10:39–47.
[15] Netten H, van Vliet LJ, Vrolijk H, Sloos WCR, Tanke HJ, Young IT. Fluorescent dot counting in interphase cell nuclei. Bioimaging 1996;4:93–106.
[16] Nychka D. Choosing a range for the amount of smoothing in nonparametric regression. J Am Stat Assoc 1991;86:653–64.
[17] Ohta Y. Knowledge-based interpretation of outdoor natural color scenes. London: Pitman; 1985.
[18] Pazzani MJ. Searching for dependencies in Bayesian classifiers. In: Fisher D, Lenz HJ, editors. Learning from data: AI and statistics V. New York: Springer-Verlag; 1996. Chapter 23, p. 239–48.
[19] Sahami M. Learning limited dependence Bayesian classifiers. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI; 1996. p. 335–8.
[20] Silverman BW. Density estimation for statistics and data analysis. New York: Chapman & Hall; 1986.