Technical Report

UCAM-CL-TR-474
ISSN 1476-2986

Number 474

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

A Bayesian methodology and probability density estimation for fluorescence in-situ hybridization signal classification

Boaz Lerner

October 1999

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

https://www.cl.cam.ac.uk/

Technical reports published by the University of Cambridge
Computer Laboratory are freely available via the Internet:

https://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Abstract

Previous research has indicated the significance of accurate classification of fluorescence in-situ hybridization (FISH) signals when images are captured in a fixed focal plane without relying on an autofocusing mechanism. Based on well-discriminating features and a trainable neural network (NN) classifier, a previous system enabled highly-accurate classification of valid signals and artifacts of two fluorophores. However, since training and optimisation of an NN require extensive resources and experimentation, we investigate in this work a simpler alternative for the NN classifier – the naive Bayesian classifier (NBC). The Bayesian methodology together with an independence assumption allow the NBC to predict the *a posteriori* probability of class membership using estimated class-conditional densities. Densities measured by three methods: single Gaussian estimation (SGE; parametric method), Gaussian mixture model (GMM; semi-parametric method) and kernel density estimation (KDE; non-parametric method) are evaluated for this purpose. The accuracy of the NBC employing data modelled by SGE is found to be similar to that based on GMM, slightly inferior to that based on KDE but widely inferior to that of the NN. Therefore, when supporting the two classifiers, the system enables a trade-off between the NN performance and the NBC simplicity. Finally, the evaluation of the NBC accuracy provides a mechanism for both model and feature selection.

1 Introduction

Fluorescence in-situ hybridization (FISH) allows the detection of specific DNA sequences in intact cells and chromosomes. It enables selective staining of various sequences in interphase nuclei and therefore the detection, analysis and quantification of specific numerical and structural chromosomal abnormalities within these nuclei.

Digital microscopy in FISH allows the application of image analysis techniques for automation of time consuming tasks, such as dot counting. Dot counting, the enumeration of signals (also called dots or spots) within the nuclei, is considered one of the most important applications of FISH. One approach to dot counting relies on an auto-focusing mechanism to select the 'clearest' image for the analysis [1, 2]. However, basing dot counting on auto-focusing can have some shortcomings [3]. Instead, it has recently been proposed [3] to base FISH dot counting on images that are sampled at a fixed focal plane. This method is motivated by the assumption that nuclei are approximately uniformly distributed in the sample, so that translations at a fixed focal plane will provide a statistically equivalent sample as projections through different focal planes. The method enables most of the shortcomings of auto-focusing to be overcome, since it shortens the length of image acquisition and requires no special instrumentation. However, since the system captures images that contain many more unfocused signals, its

ability to distinguish between focused and unfocused signals needs to be better than 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 [4], and a highly-accurate neural network (NN) classifier of these signals into real (valid) signals and artifacts, respectively [3].

Since the optimisation and training of the NN classifier require a large amount of resources and experimentation, we are investigating in the present paper a simpler, yet very powerful [5, 6], approach for FISH signal classification, namely the naive Bayesian classifier (NBC). The NBC predicts the *a posteriori* probability of class membership using Bayes' theorem, an independence assumption and the class-conditional probabilities (for discrete variables) or densities (for continuous variables) of each of its observable variables. A pattern is then assigned to the class with the highest posterior probability. If found accurate enough, the NBC may provide an attractive alternative to the NN.

Section 2 of the paper describes the procedure we use to acquire FISH images, while Section 3 depicts a methodology for multi-spectral FISH image analysis and signal measurement. Sections 4 and 5 present, respectively, the naive Bayesian classifier and three estimation methods for modelling the classifier conditional densities. Finally, Section 6 describes the experimental study and its results, while Section 7 summarises the work.

2 Biological materials and methods

2.1 Slide preparation

The interphase nuclei preparations from amniotic fluid were made using the method by Klinger *et al.* [7] with minor modifications. 1-2ml of amniotic fluid was centrifuged and the cell pellet washed in PBS warmed to 37°C. The cells were resuspended in 75mM Potassium Chloride (KCl) and put directly on to slides coated with APES (Sigma) and incubated at 37°C for 15 minutes. Evaporation of PBS was compensated with filtered distilled water. Excess fluid was carefully removed and replaced with 100ml of 3% Carnoys fixative, 70% 75mM KCl at room temperature for 5 minutes. The excess fluid was carefully removed and 5 drops of fresh fixative were dropped on to the cell area. Slides were briefly dried on a 60°C hotplate, and then either used immediately for hybridization or dehydrated through an alcohol series and stored at -20°C until required.

2.2 Hybridization

Target areas were marked on the slides using a diamond tipped scribe. Target DNA was denatured by immersing in 70% formamide:30% 2xSSC at 73°C for 5 minutes. 10 μL of probe mix containing spectrum orange LSI 21 and spectrum green LSI 13 (Vysis UK) was applied to the target area and a coverslip placed over the probe solution. Coverslips were sealed using

rubber cement and slides placed in a pre-warmed humidified container in a 37°C incubator for 16 hours. Coverslips were removed and slides washed in 0.4xSSc/0.3%NP-40 solution at 73°C for 2 minutes. Slides were then placed in 2xSSC/0.1% NP-40 solution at room temperature for 1 minute. When completely dried 10 μL of DAPI II counterstain (Vysis UK) was applied to the target area and sealed under a coverslip.

2.3 Fluorescence microscopy

Slides were screened under a Zeiss axioplan epifluorescence microscope using x100 objective. Signals were viewed using appropriate filters and images acquired using a CCD camera and SmartCapture software (Vysis UK). Slides were scanned by starting in the upper left corner of the coverslip and moving from top to bottom. Images were captured by stopping at random intervals. Red and green signals were seen on blue DAPI stained nuclei, corresponding to chromosomes 21 and 13 respectively. 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. A total of 400 images were collected from five slides and stored in TIFF format.

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 and in various combinations, image processing can be facilitated [8]. Nuclei can be analysed using the blue channel of the RGB image, whereas the signals can be analysed using the red and green channels.

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 [4]. In this work, RGB colour is recorded during the acquisition stage because pre-processing, as well as nuclei and signal segmentations, are performed more easily using this colour format than using the conventional conversion of the image to gray-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 can be easily resolved due to their dif-

ferent hues [4]. Therefore, when measuring signal features we follow Ohta [9] and convert RGB to HSI format using

$$H = \arctan 2(3^{1/2}(G - B), (2R - G - B)), \tag{1}$$

$$S = 1 - 3(min(r, g, b)) \tag{2}$$

and

$$I_1 = (R + G + B)/3, \tag{3}$$

where $r = R/(R + G + B)$, $g = G/(R + G + B)$ and $b = B/(R + G + B)$, and R, G and B are the intensities in the three channels, respectively.

Segmentation on each of the three channels of the RGB image using global thresholds yields the image nuclei and red and green signals [4]. Noise elimination and boundary smoothing of nuclei, as well as spatial correlation between nuclei and signals, complete the segmentation.

Following segmentation, signals are characterized by sets of pixel intensities. A set (signal) can include one or many members (contiguous pixels). Since the content and dimension of each set can vary dramatically from signal to signal, raw data (intensities) are not considered discriminating enough to act as features for classification. It is therefore necessary to determine a more discriminating and compact representation of the data. One representation can be derived by measuring a set of features of the signal. The features

include Area (a size measure) and Eccentricity (a shape measure), which have been previously suggested [1]. In addition, we measured a number of spectral features [4]. 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 normalized by the Average Hue. This last 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 for some kinds of artifacts (e.g. overlap of two different fluorophores) this difference is substantially large. Two additional features of the set are the two coordinates 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_1 (Eq. 3). A motivation for choosing the last three features is given elsewhere [4]. Table 1 lists and numbers the twelve features to facilitate their identification in the rest of the paper.

Different single features and sets of features have been evaluated for FISH signal classification using class-conditional probability density functions, scatter plots, a specific scatter criterion and the probability of misclassification [4]. The results of this evaluation are used in Section 6 to determine the variables, and thereby the optimal structure needed by the naive Bayesian

classifier. Then, the classifier is used to discriminate between signals of four classes: 'real red', 'artifact red', 'real green' and 'artifact green'.

4 The naive Bayesian classifier

For problems where the task is to assign test patterns to different classes, such that the probability of misclassification is minimised, the naive Bayesian classifier (NBC) provides a simple and clear method, while still enabling impressive performance. The NBC is termed naive since it makes use of a simplifying assumption that its observable variables, which represent the pattern features, are conditionally independent given the class variable. The classifier can be viewed as a special form of a Bayesian network [10], in which all the edges are directed from the class variable to the observable variables (Figure 1).

The NBC consists of a finite set $U = \{X_1, X_2, \ldots, X_m, C\} = \{\boldsymbol{X}, C\}$ of random variables, where $X_1, \ldots, X_m$ are the observable variables that represent the features, and C is the class variable with K states. The NBC assigns a test pattern $\mathbf{x}$ to the class C_k $(k = 1, \ldots, K)$ with the highest *a posteriori* probability

$$P(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)P(C_k)}{p(\mathbf{x})} \tag{4}$$

where $p(\mathbf{x}|C_k)$ is the *class-conditional* probability density, $P(C_k)$ is the *a priori* probability for class C_k, and $p(\mathbf{x})$, the *unconditional* density, normalises the

posterior probability such that $\sum_k P(C_k|\mathbf{x}) = 1$.

Using the NBC independence assumption, and omitting $p(\mathbf{x})$ which is common to all the states of the class variable, the posterior probability can be written as

$$P(C_k|\mathbf{x}) \propto p(\mathbf{X} = \mathbf{x}|C_k)P(C_k) = \prod_{i=1}^{m} p(X_i = x_i|C_k)P(C_k) \tag{5}$$

where $\mathbf{X} = \mathbf{x}$ represents the event that $X_1 = x_1 \wedge X_2 = x_2 \wedge \ldots \wedge X_m = x_m$ and $\prod_{i=1}^{m} p(X_i = x_i|C_k)$ is the likelihood for $\mathbf{x}$. Both $P(C_k)$ and $p(\mathbf{x}|C_k)$ can be estimated from the data; The estimation of $p(\mathbf{x}|C_k)$ is described in Section 5, while $P(C_k)$ is the relative frequency of patterns belonging to C_k out of all the patterns in the data.

5 Estimation of class-conditional densities

When modeling a Bayesian network such as the NBC for classification, we need to estimate $p(X_m|C_k)$, the one-dimensional class-conditional probabilities (for discrete variables) and/or the one-dimensional class-conditional probability densities (for continuous variables) for each class C_k and variable X_m. We use for this a finite number of data points $\mathbf{x}^n$, $n = 1, \ldots, N_k$, where N_k is the number of training patterns in class C_k.

Since the class-conditional densities are usually modelled by parametrized functional forms, $p(x|C_k)$ are referred to as *likelihood* functions for x. We therefore use the maximum likelihood solutions for estimating the densities

(probabilities) of the continuous (discrete) variables. For a discrete variable, this solution is given by the sample frequency for each value of the variable (that is the number of times the value is observed divided by the total number of observations). For a continuous variable, we need to estimate the density for the variable values.

The features of Section 3 include one discrete feature, Area, while all the other features are continuous. Therefore, we model the Area probability distribution using its sample frequency and estimate the probability densities for all the other features. We estimate densities using three methods that assume 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 (SGE)

In most previous work that dealt with continuous variables, data were either discretized [6] or assumed to be generated by a parametric model that is based on a single Gaussian distribution [5]. 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 can be written as

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{6}$$

where μ and σ are the mean and standard deviation of the distribution, respectively. These two parameters satisfy

$$\mu = \varepsilon[x] = \int_{-\infty}^{\infty} xp(x)\,dx \tag{7}$$

$$\sigma^2 = \varepsilon[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,dx \tag{8}$$

where $\varepsilon[\,\cdot\,]$ denotes the expectation.

The SGE for the mean and standard deviation of the normal distribution, measured for example by the maximum likelihood procedure, is [11]

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x^n \tag{9}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N} (x^n - \hat{\mu})^2 \tag{10}$$

where N is the number of training patterns (which is N_k when modelling the class-conditional densities for each class C_k). 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 (KDE)

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

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h} H\left(\frac{x - x^n}{h}\right) \tag{11}$$

using kernel functions

$$H(t) = \begin{cases} 1 & |t| < 1/2 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

with width h centred around each of the training data points x^n.

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

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{1/2}} \exp\left\{-\frac{\| x - x^n \|^2}{2h^2}\right\} \tag{13}$$

for each class-conditional density.

Non-parametric models can replace parametric models, which may not hold for some domains where data are not normally distributed. For this reason, John and Langley [5] suggested to replace SGE with KDE when modeling 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 (GMM)

Semi-parametric methods try to combine the benefits of both parametric and non-parametric methods and to provide techniques that are not restricted to specific functional forms, and yet where 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. Similarly to KDE, GMM estimates the data density using a linear combination of basis functions. 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 [11]

$$p(x) = \sum_{j=1}^{M} p(x|j)P(j), \tag{14}$$

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

Assuming the component densities are Gaussian distribution functions with means μ_j and standard deviations σ_j, and using Equation 14, the density

is modelled by

$$p(x) = \sum_{j=1}^{M} \frac{1}{(2\pi\sigma_j^2)^{1/2}} \exp\left\{-\frac{\| x - \mu_j \|^2}{2\sigma_j^2}\right\} P(j). \tag{15}$$

The difficult part in estimating data using a GMM is to find its parameters, μ_j, σ_j and $P(j)$. 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 EM algorithm [12], is employed in the experiment described in Section 6. More details on the expectation-maximization (EM) algorithm can be found in [11, 12].

6 Experimental study

To evaluate the NBC capability in classifying FISH signals, we first have to find an optimal network structure for the classifier. We can either search exhaustively within the network space for the 'best' structure, learn the structure from the data [10] or determine the structure using prior knowledge. Such knowledge [4] suggests that several feature representations are preferred for accurate classification of the signals into the four classes of Section 3. If we select the class and observable variables of the NBC to represent these classes and features, respectively, we can employ the NBC for FISH signal classification. Then, we only need to estimate the class-conditional probability densities (or the class-conditional probability for the Area feature if the

Area variable is included in the structure) for each variable given each of the four states of the class variable.

We estimate the class-conditional probability densities using the three methods described in Section 5 – SGE, KDE and the GMM. Figure 2 and Figure 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. Figures 2 and 3 demonstrate that data modelling using SGE and GMM is generally similar, where modelling using KDE is frequently spikier (especially for the artifact classes) since it depends more on the actual training data points. In the experiments, the width parameter h of KDE (Equation 13) is set to $1/\sqrt{N_k}$ (where N_k is the number of training data points in C_k) as this choice guarantees that the parameter shrinks to zero as the number of instances goes to infinity [5]. Thus, when KDE is employed for data modelling for the NBC, density estimation becomes increasingly local as the number of training points increases. Ten Gaussians are selected by the GMM to estimate the densities of each of the features. This number is considered to be reasonable based on preliminary visualisation of the data. In addition, ten iterations of the EM algorithm are found to be sufficient in order to achieve convergence of the GMM parameters.

The experiments to evaluate the classification accuracy on the training and test sets are conducted using ten-fold cross-validation (CV-10). That

is, we randomly partitioned the data into ten disjoint sets, use nine sets for training the NBC and the remaining set for the test. We then repeat this procedure ten times using all the different possible test sets and average the results. In each partition of the data, density estimation by the three methods is performed using the same training data sets.

Table 2 and Table 3 show the NBC accuracy on the test set for different single features (Table 2) and sets of features (Table 3) when the class-conditional densities are modelled using SGE, KDE and the GMM. In the tables, each single feature or set of features demonstrates a different structure of the NBC, as each feature is represented by another observable variable. The accuracy for each pattern is determined by comparing the class having the highest posterior probability for this pattern with the class selected by the cytogeneticist as the correct one. The posterior probability is the product of the likelihood and the prior probability (Equation 4). The likelihood is the product of all the individual class-conditional densities for the features (Equation 5), whereas the prior probability is given by the relative frequency of samples of each state of the class variable out of the total number of data points.

The classification accuracy of the NBC can be utilised to compare the benefits of using each of the estimation methods. The comparison, shown in Tables 2 and 3 reveals that the accuracy of the NBC is usually similar when density is estimated by the three methods with an advantage of KDE.

However, for k features, n training samples and m test samples, KDE must perform $O(n^2k)$ and $O(nmk)$ evaluations during the training and the test, respectively, where SGE, for example, needs only $O(nk)$ and $O(mk)$ evaluations, respectively.

The accuracy of the classifier can also be used to evaluate the different network structures and thereby the different features. It is demonstrated in Table 2 and Table 3 that some of the features, such as the Average (7) and Maximum Hue (6), as well as the Area (1) and Average Grey Intensity (12), enable well-discriminating representations of the signals. Furthermore, examination of the NBC performance using feature sets that contain dependent features (e.g., {1,3,7,12} and {1,6,7}) reveals that even when the NBC independence assumption is slightly violated the classifier still maintains its accuracy.

Finally, the NBC accuracy is compared in Table 2 and Table 3 to that achieved by a multilayer perceptron (MLP) neural network classifier [3, 4]. The MLP outperforms the NBC in every case and regardless of the density estimation method used by the NBC. As the number of dependent features grows in the set, this superiority increases (see e.g. feature set 'All' in Table 3). This is due to the increased violation of the independence assumption by the NBC, whereas the MLP hidden layer extracts better and better data representations for larger feature sets (even if they contain some correlated features). The MLP configuration that is tested here is the 'monolithic'

strategy of Lerner *et al.* [3], in which classification of signals into the four classes is implemented simultaneously. Other classification strategies have succeeded in classifying correctly around 86% of this FISH data [4].

7 Discussion

We study the accuracy of a special Bayesian network, naive Bayesian classifier (NBC), in classifying FISH signals. Accurate signal classification has been found essential for precise dot counting where images are captured without using an auto-focusing mechanism [3, 4].

The NBC consists of two elements. The first is the network structure, which is determined in this study using previous knowledge about the possible states of the class variable and the significance of different features to represent FISH signals. That is, variables representing the most significant features are employed as the NBC observable variables. The second component of the classifier is a set of parameters that quantifies the structure. These are used to estimate the class-conditional probability densities of each of the network observable variables given each of the states of the class variable. The densities are employed under a Bayesian methodology to derive the posterior probability for assigning an unseen signal pattern to the different classes. The pattern is then assigned to the class with the highest posterior probability among the classes.

Three approaches to 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 single Gaussian estimation (SGE) as the sample average and standard deviation, respectively. However, the particular form of the parametric function chosen might be incapable of providing a good representation of the true density. Non-parametric estimation techniques, like kernel density estimation (KDE), 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. Semi-parametric estimation methods, such as the Gaussian mixture model (GMM), 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. Following the experiments, it is also expected that by employing different (carefully-determined) numbers of Gaussians to the estimation of the different class-conditional densities, some improvement in the accuracy due to the GMM may be achieved.

The three approaches are evaluated by their accuracy to estimate density

for the NBC. They are also compared to the multilayer perceptron (MLP) neural network, another semi-parametric estimation method. The three density estimation techniques achieve comparable performance (with some advantage to KDE), which is inferior to that based on the MLP. This inferiority can be attributed to the (naive) assumption of independence of the NBC and to the MLP architecture. The MLP combines in the output layer (class variable) representations derived from hidden units that extract different nonlinear representations from the same features, and thereby enable richer and more accurate modelling of the feature space.

Accurate modelling of probability densities of patterns of high dimensionality is extremely difficult. By making use of the NBC conditional independence assumption we avoid this problem. We 'break' the high-dimensional patterns into one-dimensional (single variable) patterns and estimate separately the class-conditional densities for these patterns. Employing FISH signal features that are known to be independent from each other can therefore exploit the simplicity and the highest performance of the NBC. In other cases where features are known to be correlated to each other, the independence assumption is violated and the accuracy of the NBC may decrease. Nevertheless, in our experiments as in other experiments [5, 6] the NBC is found to maintain high classification capability even when the data present some degree of dependency. Finally, the exploitation of the classification accuracy of the NBC may be used to rank different network structures (sets of

features), and thereby to provide a methodology for feature selection.

Acknowledgements: This work was supported by EPSRC contract GR/L51072: *Automatic Analysis of FISH Images.* The author thank the members of this project for their collaboration: W.F. Clocksin, C. M. Bishop, M. Hultén and S. Dhanjal.

References

[1] H. Netten, L. J. van Vliet, H. Vrolijk, W. C. R. Sloos, H. J. Tanke, and I. T. Young. Fluorescent dot counting in interphase cell nuclei. *Bioimaging*, 4:93–106, 1996.

[2] H. Netten, I. T. Young, L. J. van Vliet, H. J. Tanke, H. Vroljik, and W. C. R. Sloos. FISH and chips: Automation of fluorescent dot counting in interphase cell nuclei. *Cytometry*, 28:1–10, 1997.

[3] B. Lerner, W. F. Clocksin, S. Dhanjal, M. A. Hultén, and C. M. Bishop. Automatic signal classification in fluorescence in-situ hybridization images. Technical Report 466, Computer Laboratory, University of Cambridge, May 1999.

[4] B. Lerner, W. F. Clocksin, S. Dhanjal, M. A. Hultén, and C. M. Bishop. Feature representation for the automatic analysis of fluorescence in-situ

hybridization images. Technical Report 464, Computer Laboratory, University of Cambridge, May 1999.

[5] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Francisco, CA: Morgan Kaufmann Publishers, 1995.

[6] R. Kohavi, B. Becker, and D. Sommerfield. Improving simple bayes. In *The 9th European Conference on Machine Learning*, 1997.

[7] K. Klinger, G. Landes, D. Shook, R Harvey, L. Lopez, P. Locke, T. Lerner, R Osathanondh, B Leverone, T. Houseal, K Pavelka, and W. Dackowski. Rapid detection of chromosome aneuploidies in uncultured amniocytes by using fluorescence in situ hybridisation (FISH). *Am. J. Hum. Genet*, 51:55–65, 1992.

[8] K. R. Castleman and B. S. White. Dot count proportion estimation in FISH specimens. *Bioimaging*, 3:88–93, 1995.

[9] Y. Ohta. *Knowledge-based Interpretation of Outdoor Natural Color Scenes.* Pitman Publishing Limited, London, 1985.

[10] D. Heckerman. A tutorial on learning with Bayesian networks. Technical report MSR-TR-95-06, Microsoft Research, March 1995.

[11] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38, 1977.

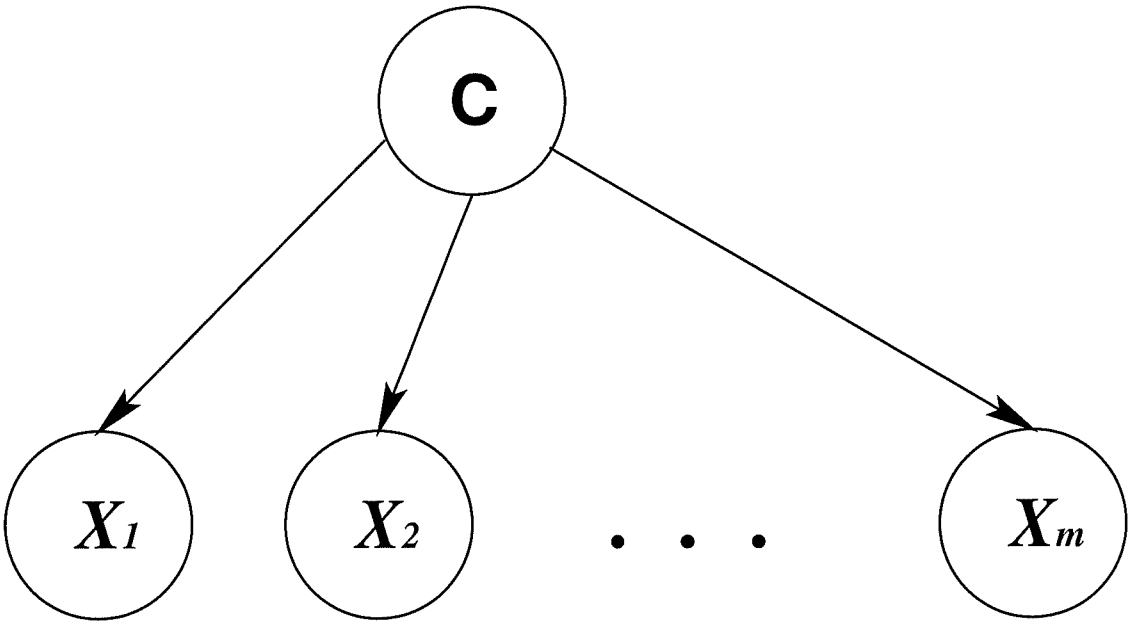

Figure 1: The naive Bayesian classifier depicted as a Bayesian network in which the observable variables (X_1,X_2, ... X_m) are conditionally independent given the class variable (C).

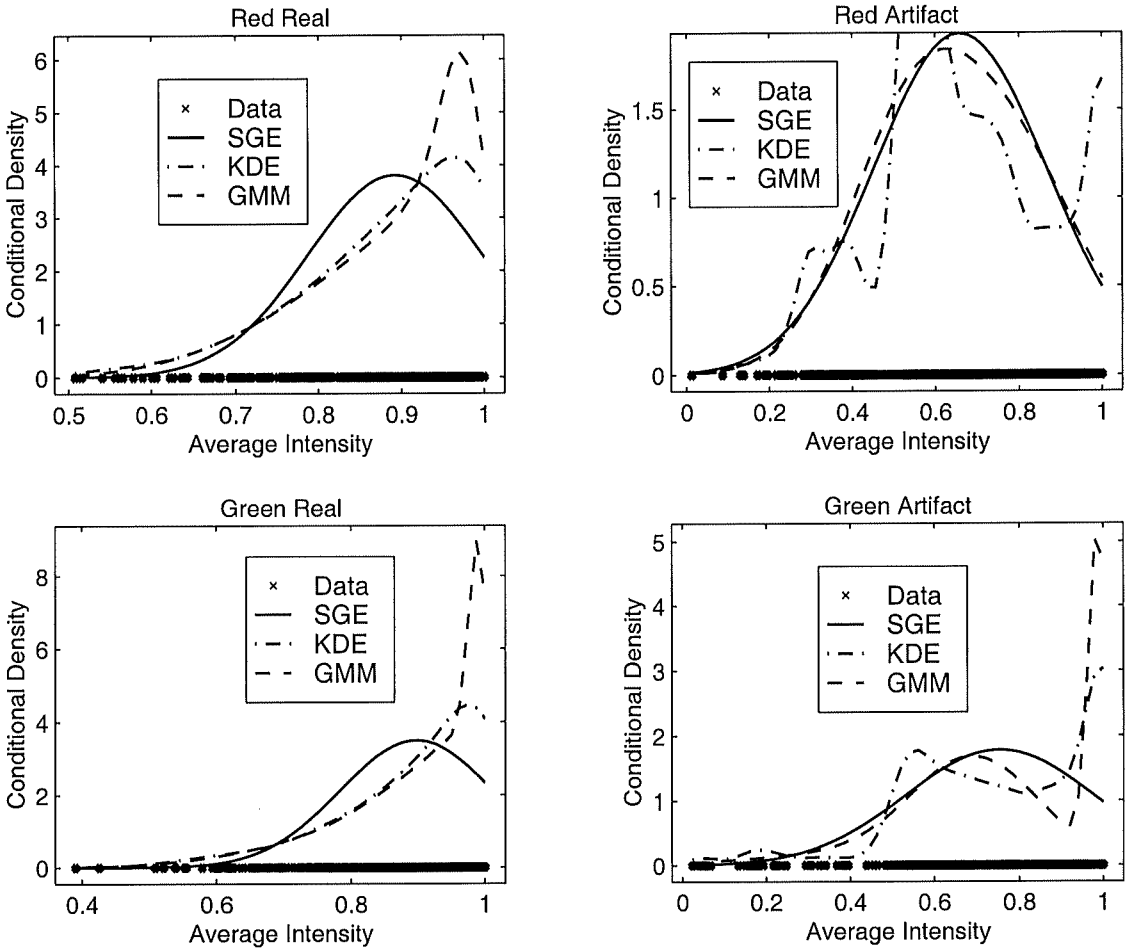

Figure 2: Class-conditional density estimation for the Average Channel Intensity feature given each of the four states of the class variable. Modelling is performed using three methods – single Gaussian estimation (SGE), kernel density estimation (KDE) and the Gaussian mixture model (GMM).

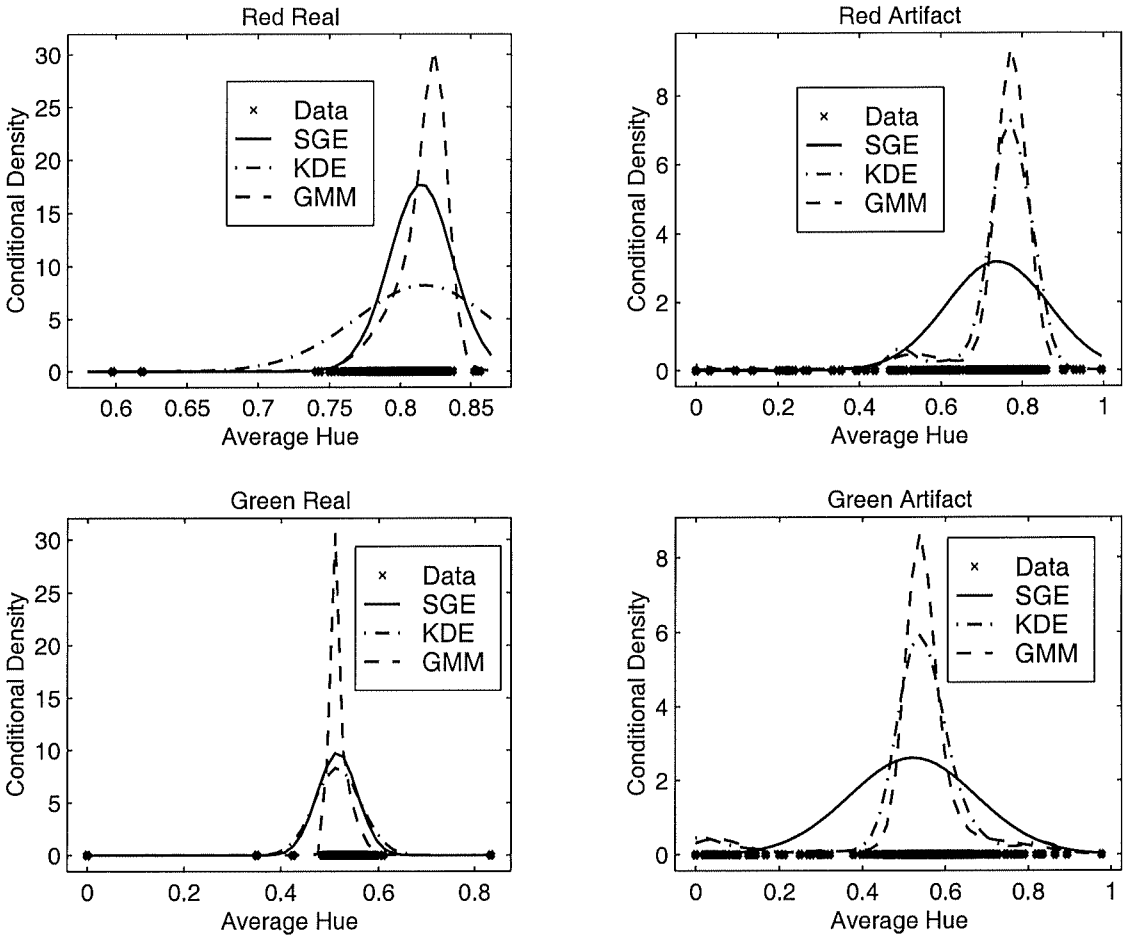

Figure 3: Class-conditional density estimation for the Average Hue feature given each of the four states of the class variable. Modelling is performed using three methods – single Gaussian estimation (SGE), kernel density estimation (KDE) and the Gaussian mixture model (GMM).

Table 1: The set of features studied in the work. 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, 2 are abbreviations for the two coordinates of the eigenvector corresponding to the largest eigenvalue of the red and green intensity components of the signal.

Number	Feature	Number	Feature
1	Area	7	Average Hue
2	Eccentricity	8	Hue Texture
3	Total Channel Intensity	9	Delta Hue
4	Average Channel Intensity	10	Eig. 1
5	Texture	11	Eig. 2
6	Maximum Hue	12	Average Grey Intensity

Table 2: The naive Bayesian classifier accuracy (%), represented by the mean and standard deviation (in brackets) when the class-conditional probability densities for single features are estimated by single Gaussian estimation (SGE), kernel density estimation (KDE) and the Gaussian mixture model (GMM). Feature numbers are defined by Table 1 and results for the 'best' seven features, are arranged by the performance using KDE from best to worst. The probability distribution (instead of density) is computed for the Area feature (1). The accuracy is compared to that achieved by a multilayer perceptron (MLP) neural network [4].

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)
11	52.3 (2.3)	52.8 (2.5)	42.4 (11.1)	53.3 (2.5)
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)

Table 3: The naive Bayesian classifier accuracy (%), represented by the mean and standard deviation (in brackets) when the class-conditional probability densities for different feature sets are estimated by single Gaussian estimation (SGE), kernel density estimation (KDE) and the Gaussian mixture model (GMM). Feature numbers are defined by Table 1 and results are arranged by the performance using KDE from best to worst. The probability distribution (instead of density) is computed for the Area feature (1). The accuracy is compared to that achieved by a multilayer perceptron (MLP) neural network [4].

Feature Combination	SGE	KDE	GMM	MLP
1, 6, 9, 12	65.0 (1.4)	79.2 (2.2)	71.1 (4.6)	82.3 (2.5)
1, 7, 12	71.8 (2.1)	79.1 (2.5)	77.1 (1.5)	82.0 (2.1)
1, 3, 7, 12	74.7 (2.3)	79.0 (2.7)	79.0 (1.6)	82.7 (2.2)
All	68.8 (3.1)	78.7 (2.5)	66.5 (5.8)	84.9 (1.7)
1, 6, 12	76.0 (1.8)	78.6 (2.6)	77.2 (2.1)	81.0 (2.3)
1, 6, 7	75.6 (2.3)	78.6 (2.3)	76.4 (2.6)	80.5 (1.8)
1, 7	75.7 (2.1)	77.9 (2.8)	78.9 (1.8)	78.0 (2.2)
1, 4, 7	72.6 (2.8)	77.7 (2.4)	76.4 (2.4)	81.3 (1.9)
1, 4, 6	73.7 (2.3)	77.5 (2.3)	75.7 (2.9)	80.8 (2.2)