On the Classification of a Small Imbalanced Cytogenetic Image Database

Boaz Lerner, Josepha Yeshaya, and Lev Koushnir

Abstract—Solving a multiclass classification task using a small imbalanced database of patterns of high dimension is difficult due to the curse-of-dimensionality and the bias of the training toward the majority classes. Such a problem has arisen while diagnosing genetic abnormalities by classifying a small database of fluorescence in situ hybridization signals of types having different frequencies of occurrence. We propose and experimentally study using the cytogenetic domain two solutions to the problem. The first is hierarchical decomposition of the classification task, where each hierarchy level is designed to tackle a simpler problem which is represented by classes that are approximately balanced. The second solution is balancing the data by up-sampling the minority classes accompanied by dimensionality reduction. Implemented by the naive Bayesian classifier or the multilayer perceptron neural network, both solutions have diminished the problem and contributed to accuracy improvement. In addition, the experiments suggest that coping with the smallness of the data is more beneficial than dealing with its imbalance.

Index Terms—Classification, dimensionality reduction, genetic diagnosis, imbalanced data, multilayer perceptron (MLP), naive Bayesian classifier (NBC), small sample size.

1 Introduction

A large error rate of a classifier is usually associated with the inherent complexity of the classification task. However, when the sample size is finite, other aspects, such as small sample size, large number of features, and the complexity of the classification rule, may also deteriorate classifier accuracy [1]. If the data are also imbalanced (or skewed), i.e., the classes have different a priori probabilities, a further decline in accuracy is expected [2], [3]. For example, if 99 percent of the data belong to one of two classes, a learning algorithm will probably fail to achieve better than the 99 percent accuracy that a trivial algorithm classifying any pattern to the majority class achieves. Moreover, the former algorithm will almost always fail on patterns of the minority class. In this study, we experimentally investigate solutions to the smallness and imbalance of the data using a small imbalanced image database used for genetic abnormality diagnosis.

One of the main methods to diagnose genetic abnormalities is fluorescence in-situ hybridization (FISH). Using FISH, various DNA sequences are stained, creating fluorescent signals that enable the detection, analysis, and quantification of numerical and structural genetic abnormalities [4], [5]. Analysis of images representing genetic numerical abnormalities is vital in clinical inspection aimed at prenatal and tumor diagnoses as well as in other applications [5]. For example, DNA sequences composing chromosome 21 in the human cell are analyzed using FISH images in order to detect an extra copy of this chromosome which indicates Down syndrome.

Current systems are successful in FISH image analysis [6], [7] and classification of dot-like FISH signals [8], [9]. However, since the conformation of the inspected sequence and, thus, of the fluorescent signal, changes during DNA replication along the cell cycle [10], non-dot-like signals are frequently found in many FISH applications and especially in clinical routine [11], [12].

In this study, we expand previous research [8], [9] in several directions. First, we identify three non-dot-like signal types that, together with the dot-like signal type and the artifact (noise) signals, define a five-class classification problem. We then develop a methodology allowing the detection and classification of signals of these types using either the naive Bayesian classifier (NBC) or the multilayer perceptron (MLP) neural network. Since the proposed methodology is general, other classifiers can be employed as well. Three density estimation paradigms are evaluated for the NBC—parametric, semiparametric, and nonparametric; each proposes a different NBC. Each of these paradigms, along with the MLP, tackles the classification problem using either a monolithic or a hierarchical training strategy.

Most of the effort in this study is directed toward improving the classification accuracy, which has deteriorated due to using a small and imbalanced database. One solution we propose is the induction of a hierarchical classifier decomposing the classification task into four two-class classification tasks; each is accomplished by employing data which is approximately balanced. A second solution we suggest and study is balancing the data by up-sampling the minority classes until reaching the number of patterns of the majority class, followed by dimensionality reduction in order to increase the ratio between the numbers of patterns and features.

- *B. Lerner and L. Koushnir are with the Department of Electrical and Computer Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel. E-mail: {boaz, koushnir}@ee.bgu.ac.il.*
- *J. Yeshaya is with the Genetic Institute, Rabin Medical Center, Beilinson Campus, Petah-Tikva 49100, Israel. E-mail: jyeshaya@clalit.org.il.*

Manuscript received 14 Mar. 2006; revised 6 Aug. 2006; accepted 21 Sept. 2006; published online 12 Jan. 2007.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-0058-0306. Digital Object Identifier no. 10.1109/TCBB.2007.070207.

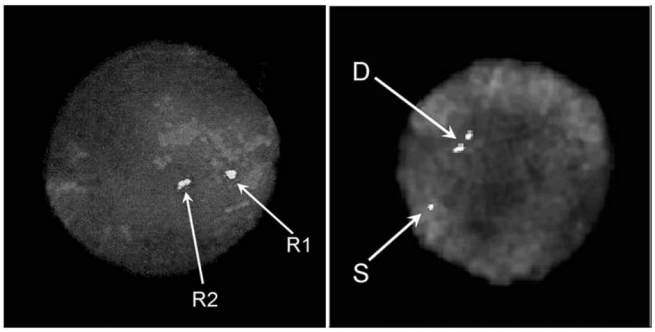


Fig. 1. Gray-level versions of FISH images showing the four types of signals associated with the replication stages. The S signal is dot-like, whereas R1, R2, and D are non-dot-like signals.

The first contribution of the paper is in the automatic classification of a small, imbalanced cytogenetic image database. We propose and examine two solutions—hierarchical task decomposition and balancing the data together with dimensionality reduction. The second contribution is in detecting and classifying non-dot-like together with dot-like FISH signals, as previous study concentrated on dot-like signals only. This ability is essential in genetic abnormality diagnosis. We begin in Section 2 by describing the cytogenetic application. In Section 3, we present FISH image analysis with emphasis on nucleus and signal segmentations. Section 4 describes a methodology of dot and non-dot-like signal classification and the building blocks of two solutions to tackle the smallness and imbalance of the data. Section 5 provides the results of applying this methodology and solutions to the cytogenetic database. We conclude the study in Section 6 with a discussion also providing some directions for future research.

2 The Cytogenetic Application

2.1 Genetic Background

When applying FISH to interphase human cells, the conformation of a fluorescent signal changes during the S-phase of the cell cycle in an ordered manner [10], [11], [12]. In the beginning of the cell cycle, the fluorescent signal appears as a single dot ("singlet"; S), representing a prereplication state. At the end of the cycle, the signal adopts a bipartite structure ("doublet"; D) representing a postreplication state. Between these two phases of the cycle appear two additional signal conformations which are more easily detected when using large probes like α-satellites. These conformation types appear as a large rounded beaded signal followed by an elongated rod-like beaded signal representing, respectively, preparation for and continuation of replication. Litmanovitch et al. [11] called these two intermediate signal conformations R1 and R2, respectively. Each of the R1, R2, and D signals is composed of one of several different settings of subsignals that define this non-dot-like signal. The S signal, however, is a dot (Fig. 1).

Usually, and especially for small probes applied to interphase cells, it is difficult to distinguish between the S and R1 signal types as well as the R2 and D types. This is the reason why the majority of work employing the FISH replication assay has dealt only with the S and D conformations by either ignoring the other two shapes or summing the S and R1 signals and the R2 and D signals into two "new" S and D entities. Nevertheless, the use of the intermediate types R1 and R2 enables raising the sensitivity (resolution) of the FISH replication method, allowing the detection of minor changes in the replication patterns of specific genes associated with genetic abnormalities [12]. Moreover, since the R1 and R2 types are detected in a significant proportion of S-phase cells, especially when exploring α-satellite sequences [11], ignoring these types in the analysis may lead to reduction of the sample size and distortion of the frequencies of occurrence of signal types. Hence, automatic classification of dot and non-dot-like FISH signals may improve the sensitivity, accuracy, and efficiency of genetic abnormality diagnosis.

2.2 Materials and Methods

Peripheral blood samples (4-5 ml whole blood) were prepared for FISH examination by regular cytogenetic methods. The samples were incubated in an RPMI medium supplemented with 5 percent fetal calf serum (FCS) and 2.5 percent phytohemaglutinin (PHA) in a $37°C$ moist chamber. After 72 hours, colchicine (final concentration of $0.1\mu g/ml$) was added to the culture for 1 hour followed by hypotonic treatment (0.075 M KCL at $37°C$ for 15 minutes) and four consecutive washes in a fresh cold fixative solution (3:1 methanol:acetic acid). The lymphocyte suspensions were stored at $-20°C$ until used. Cell suspensions were dropped on precleaned dried slides and air-dried.

We used the commercially available centromere specific probe DXZ1 (CEPX, Vysis) that consists of α-satellite sequences specific for the X chromosome and labeled with spectrum green. On each sample, a mixture of $5\mu l$ specific probe (CEP-X) was poured on the slides, covered with a 12mm circle cover glass, and sealed with rubber cement. Codenaturation was done at $76°C$ for 6 minutes, followed by incubation in a $37°C$ moist chamber for 17 hours. After hybridization, slides were washed in a salt solution ($0.4 \times SSC$) at $67°C$ for 2 minutes, followed by 1-minute wash in a second solution ($2 \times SSC/0.1\%NP40$) at room temperature in order to wash out nonspecific bounded and residual probe. The slides were allowed to dry and counterstained with $10\mu l$ of $4'$, 6-diamidino-2-phenylindole (DAPI, Vector) diluted in an antifade solution.

Slides were analyzed by an Olympus BX51 fluorescent microscope fitted with a triple band-pass filter (Chromatechnology) for coexisting detection of blue-DAPI nuclei and spectrum green signals. Simultaneously, RGB (red-green-blue) images of size 768×576 pixels were captured. The morphology of image signals was recorded by the cytogeneticist using GELFISH—a graphical environment for labeling FISH images [13], in order to provide the signals labels—S, R1, R2, D, or N (noise). These labels are required for training and evaluating the classifiers.

3 FISH Image Analysis

In the first stage of the analysis (Fig. 2), we segment isolated nuclei from their background and separate clusters of nuclei (Section 3.1). Second, we segment signals in each of the

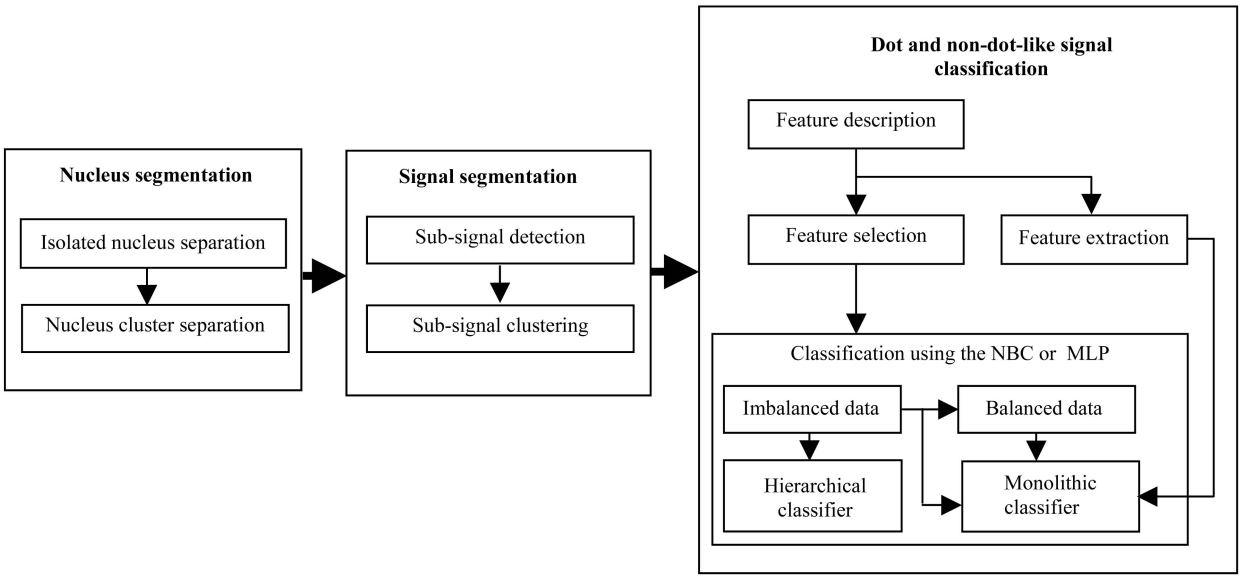


Fig. 2. A flow chart of the proposed methodology.

separated nuclei by first detecting signals as well as subsignals composing non-dot-like signals and then clustering the subsignals into signals (Section 3.2). We summarize these topics here only briefly since the paper concentrates on FISH signal classification (Section 4) rather than on image analysis.

3.1 Nucleus Segmentation

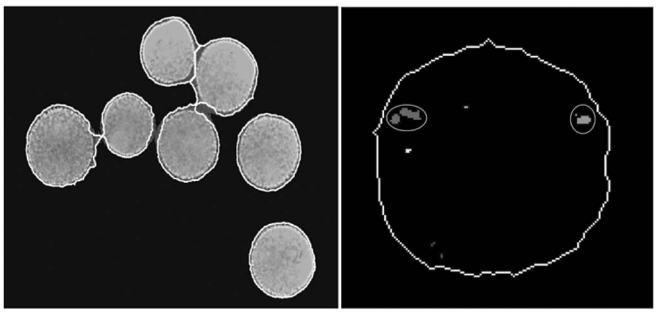
We apply two consecutive stages; each accomplishes a different nucleus segmentation objective. First, we segment isolated nuclei and clusters of nuclei from the background and, in the second stage, we separate each such cluster into the nuclei that make it. In order to segment nuclei from their background, we first eliminate image noise by a 3×3 averaging filter. We enhance image contrast by adding the top-hat filtered image to the image and then subtracting the bottom-hat filtered image [14]. Both filtered images are derived using a disk-shaped structuring element of radius 20 pixels, a radius that was found most appropriate to the task based on preliminary experiments. Then, we globally threshold the image using the Otsu method [15] in which the gray-level assuring the highest ratio of background to object variance is selected as a threshold. The result is that all isolated nuclei and nucleus clusters are separated from the background.

To separate clusters of nuclei, we apply the watershed algorithm to the binary image distance transform [16], [17]. The watershed algorithm separates connected nuclei successfully, but it tends to oversegmentation. Hence, we merge oversegmented nuclei based on their compactness (circularity).[1] First, we find in the image all objects (i.e., potential oversegmented nuclei) having compactness smaller than the average compactness computed over all objects. Second, we merge each such object to its closest object in the image if the compactness of the merged object is larger than that of either of the single objects. Following this procedure, oversegmented nuclei are correctly merged (Fig. 3a). Then, we fill small holes in the nuclei using a flood-fill operation on the background pixels, assuming the background is 4-connected [19]. Finally, we remove from the analysis small (mainly unfocused) nuclei and nuclei that are cut by the image boundaries, as the latter may contain signals in the hidden areas.

1. The compactness is defined as $C = 4\pi\frac{A}{P^2}$, where A and P are the object area and perimeter, respectively, and the maximum value $C = 1$ is reached for a circle [18].

3.2 Signal Segmentation

After nucleus segmentation, signal segmentation (Fig. 2) is performed on each separated nucleus. This is because, for numerical genetic abnormality diagnosis, we are interested in the distribution of the number of dot and non-dot-like signals per nucleus. This object-based procedure consists of the detection of signals and subsignals composing non-dot-like signals (Section 3.2.1) and clustering the subsignals into signals (Section 3.2.2). Since we study green signals (Section 2.2), the segmentation is performed on the green channel of the RGB image.



(a) (b)

Fig. 3. Examples of (a) nucleus segmentation and (b) signal segmentation on a single nucleus. Segmentation borders are marked by white contours or ellipses. Noise signals are automatically rejected and, thus, not marked.

3.2.1 Signal and Subsignal Detection

The number of pixels representing a (sub)signal in a nucleus image is much smaller than the number of nucleus (or background) pixels. A typical (sub)signal may contain 40 pixels, whereas a typical nucleus contains 10,000-20,000 pixels. Thus, the peak of the signals in the image histogram is hardly recognized and the location of an optimal threshold is difficult to determine. The result is that the conventional threshold-based segmentation methods cannot be readily applied to FISH signal segmentation and, hence, they usually require accompanying pre or postprocessing.

We preferred performing signal segmentation using a multiresolution image pyramid [20]. This pyramid is a stack of consecutive smaller replications of the image with size and resolution decreasing exponentially, as derived by decimation in factor 4. The advantage in segmenting a pyramid is mainly computational since, e.g., instead of detecting a border in the full resolution image, we use a lower resolution image for the detection and only update the border when descending from level to level until we get the full resolution image (lowest level). Each image of the pyramid is thresholded using the Otsu method [15] and the binary image of the highest level among those having the largest number of signals (i.e., the most detailed image) is projected down the pyramid consecutively to the next lower level. This is accomplished by interpolation in factor 4 (which, for a binary image, is nothing but up-sampling) and until the image containing the detected signals and subsignals is of the original size.

3.2.2 Subsignal Clustering

Following the detection of signals and subsignals, we merge subsignals into non-dot-like signals using the global k-means clustering algorithm [21] with slight changes. The algorithm starts with the subsignal most distant from the mixture mean as the first cluster (signal) center and incrementally adds as the next cluster the subsignal that, together with previous clusters, achieves the minimal sum-of-squared clustering error [22].[2] Since the number of objects to cluster is relatively small, we do not perform the k-means algorithm for each partition (i.e., we avoid repetitive recalculation of the centers), which cuts the runtime considerably. This procedure continues for increasing numbers of clusters until the number of clusters is equal to that of the subsignals, thereby providing the optimal clustering for each number of clusters. We determine the optimal partition (i.e., number of clusters) based on the maximal change in the clustering error between successive partitions. This change usually occurs immediately after the number of clusters matches the correct number of signals in the nucleus. Additional subclusters are redundant within the natural clusters, although their addition reduces the clustering error toward zero as, eventually, every subsignal may be associated with a cluster.

Finally, and since a noise signal is usually an unfocused signal having a pale narrow corona around its bright body, we screen all signals having a ratio of their average intensity after dilation to that before dilation, which is smaller than a threshold.[3] An example of signal segmentation for a specific nucleus is shown in Fig. 3b. When applied to the database described in Section 5 having 34 labeled FISH images containing 367 signals, automatic signal segmentation was 97 percent accurate compared to the segmentation of the cytogeneticist.

4 Signal Classification

In this section, we describe all aspects of our methodology of dot and non-dot-like FISH signal classification, as well as the building blocks of the solutions we suggest in order to tackle the smallness and imbalance of the data (Fig. 2). We solve a five-class classification problem for the four signal conformations representing phases in DNA replication—S, R1, R2, and D, as well as the noise (N) signals. Labels for the signals that are needed to train and evaluate the classifiers are obtained by the cytogeneticist using a graphical environment for labeling FISH images [13].

We extend previous research of dot-like FISH signal classification [8], [9] and study here the naive Bayesian classifier (NBC) and multilayer perceptron (MLP) neural network in dot and non-dot-like FISH signal classification. The NBC is modeled with class-conditional probability densities estimated using each of three types of approaches —parametric, semiparametric, and nonparametric—exemplified, respectively, by single Gaussian estimation (SGE), a Gaussian mixture model (GMM), and kernel density estimation (KDE). The MLP is configured with one layer of hidden units. The four classifiers are identified as NBC-SGE, NBC-GMM, NBC-KDE, and MLP, respectively. In addition, we explore two training strategies—monolithic[4] and hierarchical—in classifying the FISH signals by each of the four classifiers. In the first strategy, discrimination is performed by a single classifier and, in the second, discrimination is performed sequentially by specific experts decomposing the classification task. Decomposition into simpler classification tasks for which data is approximately balanced is one solution we propose to the classification of the small imbalanced cytogenetic database. Each of the decomposed tasks is a simpler task than the original task, having a smaller number of classes and a higher ratio of patterns to features; hence, the curse-of-dimensionality is diminished and the classification accuracy is expected to increase. As both majority and minority classes in the decomposed tasks are more balanced than for the original task, they can be represented and classified more accurately.

We first suggest a feature selection method to choose a well-discriminated subset from the feature set representing the FISH signals (Section 4.1). Then, we briefly introduce the NBC (Section 4.2) and MLP (Section 4.4) employed for signal classification. In between, we review three methods of density estimation for the NBC (Section 4.3). Finally, we address data balancing by up-sampling as part of our second remedy to the problem (Section 4.5).

4.1 Feature Selection

We measure a set of signal features based on [8] ("Feature description" in Fig. 2). The features include size (e.g., area

2. This is the sum of the squared euclidean distances between each subsignal center and the center of the cluster that contains the subsignals.

3. Based on experimentation, the threshold was determined to be 0.7.

4. We follow the convention and use the term "monolithic" classifier whenever it is compared to another classifier decomposing the task, e.g., hierarchically.

and different ratios of signal-related areas), shape (e.g., eccentricity and axis lengths of the bounding ellipse), hue, and intensity (measured in the RGB green plane). Other features are based on the shape descriptors (e.g., convexity and compactness) of [23].

In order to diminish the feature statistical dependence and curse-of-dimensionality, we apply feature selection. This requires a procedure to search candidate feature subsets and a criterion to evaluate each such subset [24]. We use the classification accuracy, whether of the NBC or MLP, as the criterion. Since exhaustive search is an impractical procedure even for moderate feature sets and subsets, we propose a greedy search algorithm based on the sequential forward selection procedure [24]. The algorithm starts with an empty feature subset, i.e., the first "current" subset. It evaluates every feature, finds the one having the highest value of the criterion, and adds it to the current subset. All remaining features are kept aside. In each iteration and until no features remain aside, the algorithm evaluates which of the remaining features when combined with the current subset provides the highest criterion value. This feature is added to the current subset and simultaneously excluded from the remaining features. When no features remain, we select the feature subset achieving the highest value of the criterion from among all subsets, each having the highest value of the criterion for a specific number of features.

4.2 The Naive Bayesian Classifier

The NBC [25] is a model for a finite set of random variables $\boldsymbol{U} = \{X_1, X_2, \ldots, X_m, C\} = \{\boldsymbol{X}, C\}$, where $X_1, X_2, \ldots, X_m$ are the observable variables that represent the features and C is the class variable having L states (for L classes). Albeit assuming naively that all of the observable variables are conditionally independent given the class variable, the NBC often classifies patterns accurately compared to other state-of-the-art classifiers [26]. The NBC assigns a test pattern $\boldsymbol{x}$ to the class C_K having the highest a posteriori probability

$$C_K = \arg\max_{k=1,L}\{P(C_k|\boldsymbol{x})\} = \arg\max_{k=1,L}\left\{\frac{p(\boldsymbol{x}|C_k)P(C_k)}{p(\boldsymbol{x})}\right\}, \quad (1)$$

where $p(\boldsymbol{x}|C_k)$ is the class-conditional probability (for a discrete variable) or probability density (for a continuous variable), $P(C_k)$ is the a priori probability of class C_k, and $p(\boldsymbol{x})$ is the unconditional density normalizing the product of the former two terms such that $\sum_k P(C_k|\boldsymbol{x}) = 1$. Using the NBC independence assumption and omitting $p(\boldsymbol{x})$, which is common to all states of the class variable, the posterior probability (1) can be written as

$$P(C_k|\boldsymbol{x}) \propto p(\boldsymbol{X} = \boldsymbol{x}|C_k)P(C_k) = P(C_k)\prod_{i=1}^{m} p(X_i = x_i|C_k), \quad (2)$$

where $\boldsymbol{X} = \boldsymbol{x}$ is the assignment of a state to each variable of $\boldsymbol{X}$. Assuming that all variables are continuous, $\prod_{i=1}^{m} p(X_i = x_i|C_k)$ is a product of one-dimensional (1D) class-conditional densities (thus, from now on, we will use x instead of $\boldsymbol{x}$). Both $P(C_k)$ and $p(X_i|C_k)$ can be estimated from the training data; $P(C_k)$ is the relative frequency of patterns belonging to C_k out of all of the patterns and $p(X_i|C_k)$ is usually estimated by either of the three methods described in Section 4.3.

4.3 Estimation of Class-Conditional Probability Densities

Decomposition of the computation of the class-conditional density of the NBC using (2) reduces the curse-of-dimensionality since this computation requires only linearly rather than exponentially increasing (with the dimension) numbers of patterns. We estimate $p(X_i|C_k)$ for each class C_k and variable X_i employing a training set of patterns x^n, where n gets values for each of the N_k training patterns of class C_k.

The class-conditional probability density may be estimated assuming different data generation mechanisms. In this study, we explore three density estimation methods assuming different mechanisms of data generation. Single Gaussian estimation (Section 4.3.1) assumes the data are generated from a single Gaussain distribution. Kernel density estimation (Section 4.3.2) models the data using a linear combination of kernels, each of which is located around a training pattern. A Gaussian mixture model (Section 4.3.3) estimates the data using a few Gaussians having adaptable parameters. All of these methods are only briefly summarized here, but are detailed in [22], [27], [28].

4.3.1 Single Gaussian Estimation

Usually, each 1D class-conditional density of the NBC is assumed to be Gaussian. Then, when estimated using maximum likelihood, the Gaussian mean and standard deviation are the sample average and standard deviation, respectively, leading to single Gaussian estimation (SGE).

4.3.2 Kernel Density Estimation

Nonparametric methods make no assumptions about the density functional form, but use the data to estimate the probability density. Kernel density estimation (KDE), a leading nonparametric estimation method, models the 1D density using Gaussian kernel functions [28]

$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{1/2}}\exp\left\{-\frac{\|x - x^n\|^2}{2h^2}\right\}, \quad (3)$$

where a Gaussian kernel having a width parameter h is centered around each of the N training patterns x^n. Usually, h is modeled using a parametric form such as $h = TN^{\alpha}$, where $T > 0$ is a multiplicative factor, $-1 < \alpha < 0$, and $N = N_k$ is the number of patterns in class C_k. Choosing $\alpha = -1/2$ [22] guarantees that the parameter h shrinks to zero as the number of patterns goes to infinity and, hence, KDE becomes increasingly local with the number of training patterns.

4.3.3 A Gaussian Mixture Model

Semiparametric methods are not restricted to specific functional forms (as in parametric methods) and yet the model size depends only on the problem complexity and not on the data size (as in nonparametric methods). A GMM is a semiparametric method that estimates the density using

a linear combination of $M < N$ 1D Gaussian densities $p(x|j)$ that are each weighted by a mixing coefficient $P(j)$, which is the prior probability that the jth density has generated the mixture density. The mixture density is [27]

$$\begin{aligned} p(x) &= \sum_{j=1}^{M} p(x|j)P(j) \\ &= \sum_{j=1}^{M} \frac{1}{\left(2\pi\sigma_j^2\right)^{1/2}} \exp\left\{-\frac{\|x-\mu_j\|^2}{2\sigma_j^2}\right\} P(j), \end{aligned} \tag{4}$$

where $P(j)$ satisfy the probability constraints

$$\left(\sum_{j=1}^{M} P(j) = 1, 0 \le P(j) \le 1 \ \ \forall j\right)$$

and the Gaussian densities $p(x|j)$ have means μ_j and spherical covariance matrices with standard deviations σ_j. Note that, in addition to estimating μ_j and σ_j, we should also estimate $P(j)$. Most of the methods for determining these parameters from the data are based on maximum likelihood utilizing the expectation-maximization (EM) algorithm [29].

4.4 Multilayer Perceptron Neural Network

When acting as a classifier, the MLP approximates the a posteriori probability of class membership by minimizing the error between the network output and desired target (label) [27]. The network hidden layer maps the feature space onto the target space and thereby performs internal feature extraction that alleviates the classification task and increases its accuracy. The total input to the kth unit in a layer of the MLP is

$$s_k = \sum_j w_{jk} y_j + \theta_k, \tag{5}$$

where w_{jk} is the weight connecting this unit to the jth unit of the preceding layer, the latter having an output y_j, and θ_k is the kth unit bias. Usually, the activation function that maps this unit input (5) to an output is the logistic sigmoid

$$y_k = F(s_k) = \frac{1}{1+e^{-S_k}}. \tag{6}$$

It is a differentiable function permitting the application of gradient descent-driven optimization algorithms in minimizing the output error. Also, this function has values in [0, 1] allowing the interpretation of the network outputs as class a posteriori probabilities [27].

Learning is repeated iteratively on the training set, so weights between units of different layers can be adjusted in order to minimize network error. In this study, the MLP is configured with one layer of hidden units trained by the scaled conjugate gradient algorithm [27].

4.5 Data Balancing

Whenever a class in a classification task is underrepresented (i.e., has a lower prior probability) compared to other classes, we consider the data as imbalanced [2], [3]. The main problem in imbalanced data is that the majority classes that are represented by large numbers of patterns rule the classifier decision boundaries at the expense of the minority classes that are represented by small numbers of patterns. This leads to high and low accuracies in classifying the majority and minority classes, respectively, which do not necessarily reflect the true difficulty in classifying these classes.

Most common solutions to this problem balance the number of patterns in the minority or majority classes. Either way, balancing the data has been found to alleviate the problem of imbalanced data and enhance accuracy [2], [3], [30]. Data balancing is performed by, e.g., up-sampling patterns of minority classes either randomly or from areas close to the decision boundaries. Interestingly, random up-sampling is found comparable to more sophisticated up-sampling methods [30]. Alternatively, down-sampling is performed on majority classes either randomly or from areas far away from the decision boundaries. We note that random down-sampling may remove significant patterns and random up-sampling may lead to overfitting, so random sampling should be performed with care. We also note that, usually, up-sampling of minority classes is more accurate than down-sampling of majority classes [30]. Indeed, in Section 5, we apply up-sampling as one means for balancing the cytogenetic database.

5 Experiments and Results

In this section, we experimentally study dot and non-dot-like signal classification in FISH images. We apply the methodologies of Section 4 to a cytogenetic database of 34 labeled FISH images containing 367 signals. Based on labels provided by a cytogeneticist and the taxonomy of Section 2.1, 118 signals are considered as S, 106 as R1, 43 as R2, 44 as D, and 56 signals as noise (N). These images were captured and the signals were labeled in a previous project and they are used here as is. This is the only image database available to us (and, to the best of our knowledge, the only one that exists) for studying non-dot-like FISH signals. Hence, we have had to cope with a small and imbalanced database that is labeled by a single cytogeneticist.

As a reference, we test the accuracy of the monolithic training strategy on the imbalanced data (Section 5.1.1). Since the cytogenetic database is small and heavily imbalanced (e.g., S versus D, R1 versus R2), we concentrate on two solutions to increase this accuracy. First, we apply the hierarchical strategy using specific expert classifiers in each hierarchy level (Section 5.1.2). These experts are based on the NBC (Section 4.2) or MLP classifier (Section 4.4). The hierarchical strategy alleviates the smallness of the data by decomposing the classification task and inducing simpler expert classifiers each for a smaller number of classes that are usually represented by more patterns. The design of the experts for equally populated classes as possible also counterweighs data imbalance. In the second solution (Section 5.2), we balance the data by up-sampling patterns of the minority classes until class sizes match that of the majority class (Section 4.5). We measure the effectiveness of balancing the data when the NBC and MLP are trained using the monolithic strategy (Section 5.2.1). Then, we study dimensionality reduction in lessening the smallness problem of the already balanced data (Section 5.2.2).

TABLE 1
Average Accuracies (and Std) of the Four Studied Classifiers Operating Each with Its Own Optimal Feature Subset and Parameters Using the Monolithic Strategy and Imbalanced Data

Classifier	Training accuracy (%)	Test accuracy (%)
NBC-GMM	75.0 (1.4)	67.9 (2.1)
NBC-SGE	69.5 (2.2)	67.8 (2.0)
NBC-KDE	74.6 (2.0)	65.5 (1.7)
MLP	78.5 (1.8)	67.5 (3.3)

5.1 Results for the Imbalanced Data

5.1.1 The Monolithic Strategy

We create five random replications of the data and perform, using each replication, a twofold cross-validation (CV) experiment, i.e., divide the data into two equal sets (folds), train on one set, and test on the other set before changing the roles of the sets. The reported classification accuracy is the average over the 10 folds (two folds for each of the five replications) in a test called 5x2cv. This test has small type I and type II errors when comparing classifier accuracies on small data [31]. We avoid dividing the data further in order to have an additional separate set for validation of parameters since it may undermine achieving a reasonably accurate probability estimation due to the smallness of the data. This division is especially unnecessary for the NBC, which estimates only a small number of parameters, so overfitting the data is unlikely.

Each of the NBC-SGE, NBC-GMM, NBC-KDE, and MLP classifiers is evaluated using the 5x2cv test on subsets of increasing numbers of optimal features and the subset that provides peak accuracy is determined (Section 4.1). Using this subset, model parameters leading to the highest accuracy are determined. Utilizing this methodology, we select, for the NBC-GMM, six Gaussians in order to model the probability density for each of its six optimal features and update the density parameters using a single iteration of the EM algorithm. The best generalization capability of the NBC-KDE is obtained for a multiplicative factor (Section 4.3.2) of $T = 2$ and a subset of seven optimal features. The NBC-SGE needs a subset of nine optimal features. The most suitable configuration of the MLP is of 10 hidden units in the single hidden layer and seven input units corresponding to the seven optimal features for this classifier. Training the MLP is stopped when the difference in the mean-squared error between two consecutive epochs is 10e-4 or less. We note that all four classifiers select some optimal features in common. These are the diameter of a circle having the signal area, length of the minor axis of the ellipse that bounds the signal, ratio of the major axis to the minor axis of this ellipse and the compactness (i.e., the ratio of the signal squared perimeter to area). The remaining features that are selected by each of the classifiers are based on other shape and hue features (Section 4.1).

Table 1 shows that all classifiers achieve similar accuracies using their optimal sets of features and parameters. Analyzing the classifier confusion matrices provides some explanations. For example, the confusion matrix of the NBC-KDE (Table 2) shows that 49.2 percent and 67.8 percent, respectively, of the patterns of the minority classes R2 and D are wrongly classified and the variances in their classification accuracies are relatively high. There are two main reasons for the high error rate of minority classes, as has been noted before [31]. First is that the classifier accuracy depends on the class prior probability and this probability is lower for a minority class than for a majority class. Second, having a small number of training patterns, a minority class is represented inadequately and, hence, also classified inaccurately compared to a majority class. For the same reason, a minority class may also be represented differently in the training and test sets, which causes relatively low prediction accuracy and high variance of this accuracy. The D (R2) signals are wrongly classified as R2 (D) due to similar feature values. These signals are also wrongly classified as R1 (a majority class) due to the latter numerical dominance in shaping the classes' decision boundaries. That is, the smallness and imbalance of the data undermine successful classification of minor classes and thus also weaken the classifier average accuracy.

TABLE 2
The NBC-KDE Confusion Matrix (Accuracy (Percentage) Mean and Std) for the Monolithic Strategy and Imbalanced Data

Human System	S	R1	R2	D	N
S	79.4 (3.9)	14.2 (2.8)	0 (0)	0.5 (0.9)	5.9 (3.4)
R1	18.9 (4.3)	69.0 (5.3)	5.6 (2.0)	6.5 (2.1)	0 (0)
R2	3.0 (2.6)	27.7 (3.2)	50.8 (7.7)	18.5 (8.2)	0 (0)
D	16.1 (5.2)	35.8 (4.5)	15.9 (5.3)	32.2 (8.9)	0 (0)
N	25.0 (8.9)	2.3 (2.1)	1.1 (1.8)	3.9 (2.5)	67.7 (8)

5.1.2 The Hierarchical Strategy

In the hierarchical strategy, signals are classified in each hierarchy level to either one of two classes. We first classify signals as belonging to either {S, N} or {R1, R2, D} as the noise signals are mostly similar to the S signals (Table 2) and the R1, R2, and D signals are similar to each other. In the second hierarchy level, one expert classifier distinguishes between the S and N signals and another expert classifier between the R1 and {R2, D} signals as the R2 and D signals are similar and belong to two equally sized minority classes. In the last hierarchy level, a fourth expert discriminates between the R2 and D signals, altogether accomplishing the five-class classification task. Such decomposition of the task reduces both the smallness and imbalance problems of the database. Each expert faces a reduced two-class classification task with a relatively larger and more balanced data set. These tasks are discriminating 174 signals of the {S, N} class from 193 signals of the {R1, R2, D} class, 118 signals of the S class from 56 of the N class, 106 signals of the R1 class from 87 signals of the {R2, D} class, and 43 signals of the R2 class from 44 signals of the D class.

Table 3 shows the accuracies of the NBCs and MLP trained hierarchically and after each expert is optimized to the features and parameters. The accuracies of the classifiers are similar and the only difference is that the MLP needs only one to three features, depending on the hierarchy level, whereas the other classifiers require four or more features. Not shown are the accuracies of the experts in each of the

TABLE 3
Average Accuracies (and Std) of the Four Studied Classifiers Operating Each with Its Own Optimal Feature Subset and Parameters Using the Hierarchical Strategy and Imbalanced Data

Classifier	Training accuracy (%)	Test accuracy (%)
NBC-GMM	74.1 (2.3)	69.6 (2.4)
NBC-SGE	71.3 (2.2)	69.0 (3.2)
NBC-KDE	76.9 (3.0)	66.3 (2.6)
MLP	74.1 (2.2)	68.4 (2.9)

hierarchy levels but only the overall accuracy. Generally, classification in the first hierarchy level and between S and N is more accurate (~ 90 percent) than when classifying patterns of the R1, R2, and D signals ($\sim$ 70-80 percent), again because the classification of signals of the small minority classes is very difficult. From analyzing the confusion matrices (not shown) for the different hierarchy levels, we notice that the error rates in all cases, except when distinguishing between the S and N signals, are symmetric between the classes. This corresponds to relatively data-balanced tasks (except for the S versus N task) which enables the hierarchical classifier to cope with the imbalance in the data. Finally, Table 3 shows that the hierarchical strategy improves on the monolithic strategy test accuracy (Table 1) in about 1-2 percent.

5.2 Results for the Balanced Data

Using the second suggested solution, we balance the data by randomly duplicating patterns of the classes up to the number of patterns of the largest class, S (i.e., the up-sampling method). Then, we evaluate the four classifiers trained using the monolithic strategy in discriminating FISH signals (Section 5.2.1). Also, we reduce dimensionality by extracting features from the optimal feature subset, thereby tackling the problem of the smallness of the data by improving the ratio of the number of patterns to that of the features (Section 5.2.2).

5.2.1 The Monolithic Strategy

We repeat the procedure performed for the imbalanced data for the balanced data. Fig. 4 demonstrates that the classifier test accuracies increase with the number of discriminative features selected until a certain point, after which they deteriorate due to the curse-of-dimensionality. This point determines the optimal feature subset for each classifier. The optimal feature subset selected for each classifier is similar to that selected for the imbalanced data (Section 5.1.1). Fig. 4 also shows that the NBC-KDE is less sensitive to the curse-of-dimensionality than the other classifiers, as exemplified in retaining accuracies close to the ultimate accuracy for a broad range of sizes of feature subsets. In contrast, the NBC-GMM is the most sensitive to the curse-of-dimensionality, losing its accuracy for large feature subsets.

Using its own optimal feature subset, we determine for each classifier the optimal parameters. Fig. 5a and Fig. 5b show, respectively, that the NBC-GMM test accuracy

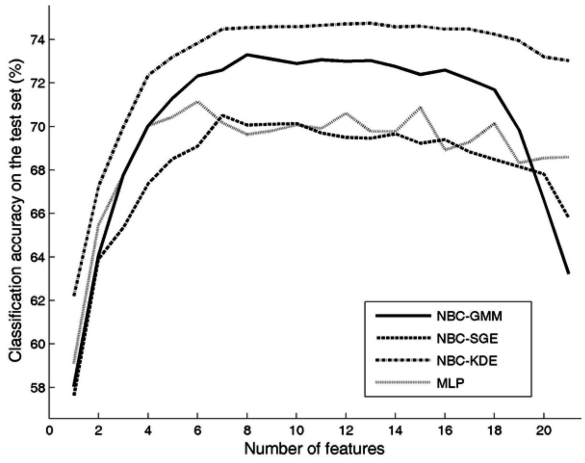


Fig. 4. The test classification accuracy for increasing numbers of optimal features for the NBC-GMM, NBC-SGE, NBC-KDE, and MLP, all trained using the monolithic strategy and balanced data.

increases with the numbers of Gaussian components and EM iterations until a certain point, after which it remains almost steady. Based on these figures and in order to achieve the highest accuracy, we select a model having 10 Gaussians which is trained for nine iterations of the EM algorithm. Similarly, and on the basis of Fig. 5c, we choose the multiplicative factor T^2 determining the Gaussian width h in (3) to be one (i.e., a Gaussian width that decreases with the square root of the number of patterns). As the MLP reaches its ultimate accuracy when using 11 hidden units (Fig. 5d), we choose this value for the MLP. Training the MLP continues until its mean-squared-error is not changed in more than 10e-4 between epochs.

Table 4 summarizes the classification accuracies of the four models, each utilizing its own optimal feature subsets and parameters. The accuracies are similar, with some advantage to the NBC-KDE. In order to decide whether this advantage is statistically significant and similarly for any two classifiers to be compared, we assume the two classifiers have the same error rate (i.e., the null hypothesis). Then, we compute a statistic from the errors of the two classifiers measured on the test set and, if our assumption holds, this statistic should obey a certain distribution. If the statistic has a probability large enough of being drawn from this distribution, we accept the hypothesis; otherwise, we reject it, saying the two classifiers have different error rates. The statistic we use, having low type I and type II errors, was suggested for the 5x2cv test in [31] before being modified in [32]. We first define $p_i^{(j)}$ as the difference between the error rates of the two classifiers on fold $j = 1, 2$ of replication $i = 1, \ldots, 5$. The average on replication i of the two folds is $\bar{p}_i = (p_i^{(1)} + p_i^{(2)})/2$ and the estimated variance is $s_i^2 = (p_i^{(1)} - \bar{p}_i)^2 + (p_i^{(2)} - \bar{p}_i)^2$. We then define a statistic that is approximately F distributed with 10 and 5 degrees of freedom [32],

$$f = \frac{\sum_{i=1}^{5} \sum_{j=1}^{2} \left(p_i^{(j)}\right)^2}{2\sqrt{\sum_{i=1}^{5} s_i^2}}. \quad (7)$$

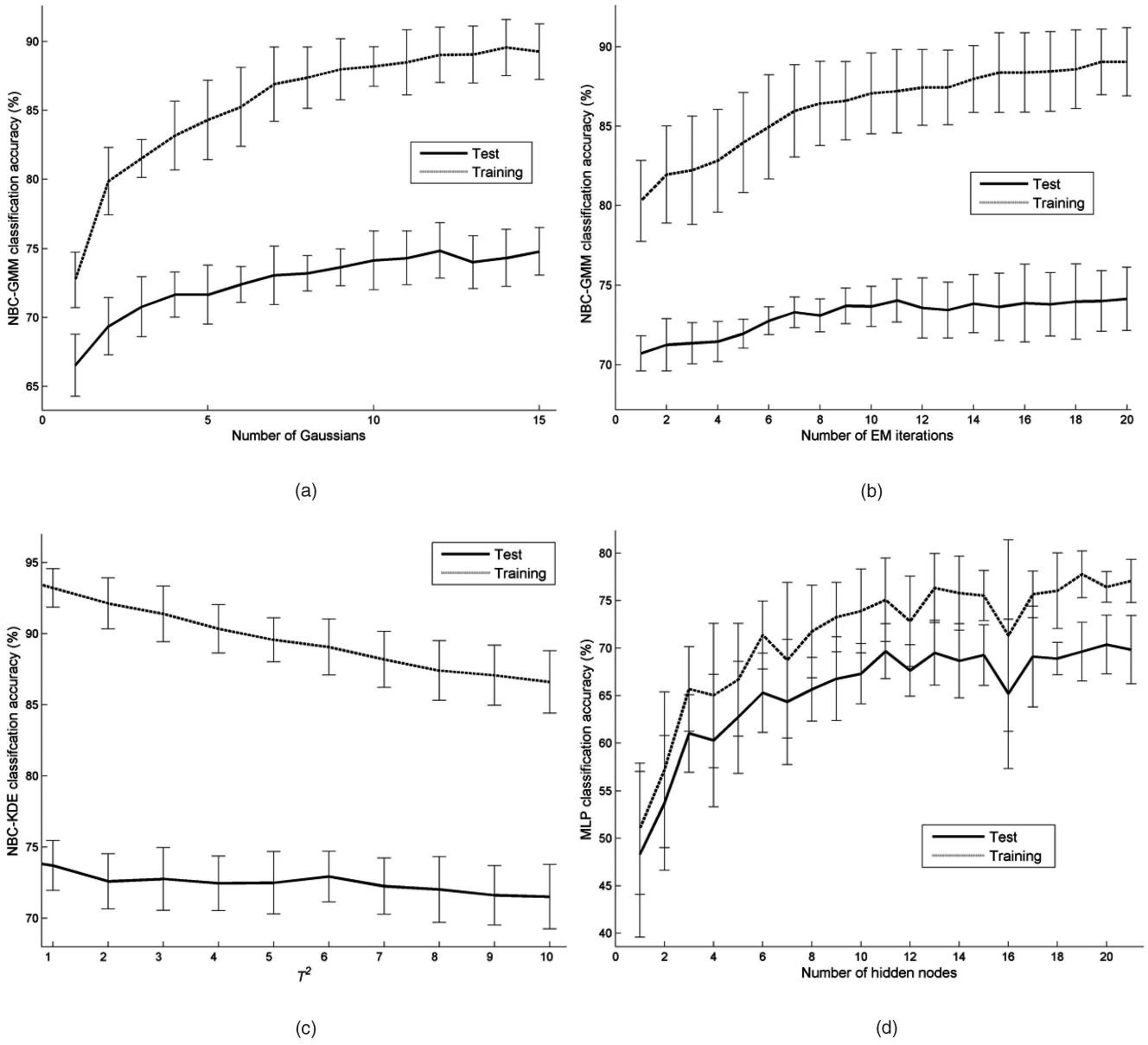


Fig. 5. The classification accuracy for the monolithic strategy, balanced data, and optimal feature subset of the (a) NBC-GMM for increasing numbers of Gaussians, (b) NBC-GMM for increasing numbers of EM iterations, (c) NBC-KDE for increasing values of the multiplicative factor determining the Gaussian width, and (d) MLP for increasing numbers of hidden nodes.

We reject the null hypothesis, i.e., declare that the two classifiers are different in accuracy with 0.95 confidence, if f is greater than 4.735. Returning to Table 4 and computing f for any pair of classifiers, we can summarize that the differences between the classifiers with respect to the above statistic are not statistically significant.

TABLE 4
Average Accuracies (and Std) of the Four Studied Classifiers Each Operating with Its Own Optimal Feature Subset and Parameters Using the Monolithic Strategy and the Balanced Data

Classifier	Training accuracy (%)	Test accuracy (%)	Size of feature set
NBC-GMM	85.9 (2.1)	73.3 (1.6)	8
NBC-SGE	79.0 (1.1)	70.5 (2.5)	7
NBC-KDE	92.4 (1.2)	74.7 (2.5)	13
MLP	77.7 (3.4)	71.1 (1.8)	6

Comparing Table 1 and Table 4 draws several interesting conclusions. First is that replicating patterns of the minority classes increases the overall training classification accuracy of the NBC-based methods as the estimation of the densities for these classes has been improved. Second, since after data balancing, the training and test sets share patterns in common, the improvement in the training classification accuracy is also reflected in the test accuracy, but the two accuracies are not independent anymore. Third, the classifiers are affected differently from balancing the data. Nonparametric classifiers gain more from balancing since they depend on the data more than parametric classifiers. For example, following data balancing, the NBC-KDE

TABLE 5
The NBC-KDE Confusion Matrix (Accuracy (Percentage) Mean and Std) for the Monolithic Strategy and Balanced Data

Human System	S	R1	R2	D	N
S	74.3 (6.6)	11.8 (4.2)	0 (0)	4.5 (2.8)	9.4 (4.7)
R1	20.7 (6.2)	50.7 (6.6)	14.3 (5.2)	16.9 (2.6)	0.9 (0.9)
R2	1.8 (2.4)	8.5 (8.4)	85.3 (8.9)	7.6 (3.9)	1.3 (1.8)
D	5.9 (2.6)	7.7 (3.2)	9.3 (4.8)	79.6 (6.9)	0.8 (0.9)
N	13.5 (5.6)	2.1 (2.5)	1.0 (1.2)	2.0 (1.6)	83.3 (4.1)

increases its test accuracy by 9.2 percent compared to NBC-SGE, which gains only 2.7 percent. The MLP, although not the most accurate classifier, is the most reliable classifier. It increases its test accuracy (similarly to the NBC-SGE and NBC-GMM but less compared to the NBC-KDE) but without overfitting the training set as the NBC-based methods do. It also uses a smaller number of features than the NBCs. This indicates both the MLP superior capability of generalization and estimation of the true classification accuracy for the cytogenetic balanced data.

We also analyze the classifier confusion matrices. For example, we compare the NBC-KDE accuracy for the balanced data (Table 5) to that for the imbalanced data (Table 2). The accuracy for the minority classes has significantly intensified (e.g., the accuracy for the R2 class has increased from ~ 50 to ~ 85 percent and that for the D class from ~ 32 to ~ 80 percent) at the expense of accuracy for the majority classes (e.g., the accuracy for the R1 class has reduced from 69 to ~ 51 percent). This pattern of improving the accuracy on the minority classes (R2 and D) at the expense of deteriorating the accuracy on the majority classes (S and R1) with an overall increase of accuracy is also found for the other classifiers.

5.2.2 Dimensionality Reduction

Besides data balancing, we also tackle the smallness of the data using feature extraction. In feature selection, we choose features from the original set based on a criterion (Section 4.1), whereas in feature extraction, we project the features onto another space and use a smaller number of feature projections rather than the features themselves (see [24] for details). Note, however, that we may apply feature extraction to the result of feature selection. The application of feature extraction allows us to combine two different remedies to the curse-of-dimensionality caused by the smallness of the data. Both data balancing and dimensionality reduction are well-known methods that were used in such cases before, but, to the best of our knowledge, they were not employed together and certainly not in FISH signal classification. By reducing the dimensionality, feature extraction improves the ratio between the numbers of patterns and features, thereby alleviating the curse-of-dimensionality.

We apply principal component analysis (PCA) [24] to the balanced data. PCA is one of the simplest and most popular feature extraction methods. Fig. 6 shows the improvement in accuracy of the MLP classifier for increasing numbers of eigenfeatures derived by projecting the cytogenetic database on increasing numbers of eigenvectors corresponding to the largest eigenvalues. Utilizing 18 eigenfeatures having less correlation among them compared to the original features led

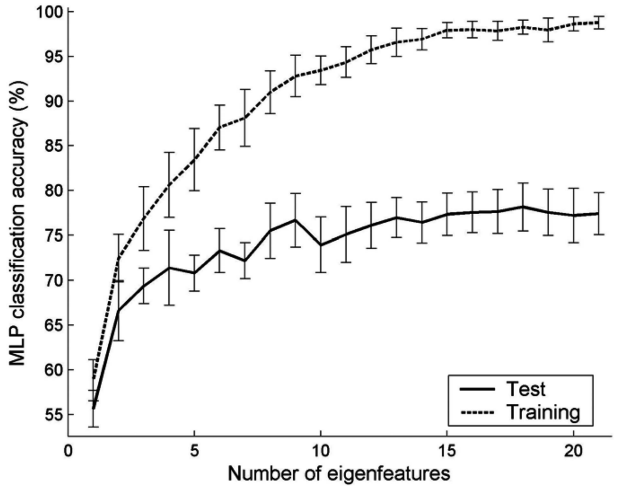


Fig. 6. The MLP classification accuracy for increasing numbers of eigenfeatures extracted from the balanced data and using the monolithic strategy.

to test accuracy of 78.1 percent, higher than that previously achieved by the MLP or any other classifier for the studied database. That is, the combination of the nonparametric MLP classifier, data balancing, and feature dimensionality reduction is most beneficial in classifying the small imbalanced FISH image database.

6 Discussion

Solving a multiclass classification task using a small imbalanced database of multivariate feature representation of patterns is complex for two main reasons. First is the curse-of-dimensionality requiring exponentially increasing numbers of patterns with the dimensionality in order to obtain accurate parameter estimation. Second is that imbalanced data leads to biased training in which the majority classes rule the classifier decision boundaries and, thus, patterns of these classes are mainly classified correctly, whereas the minority classes are overlooked and their patterns are frequently misclassified.

Since the cytogenetic database is small and heavily imbalanced, the minority classes were generally represented inadequately and differently compared to the majority classes and also between the training and test sets. This caused relatively low prediction accuracy and high variance of this accuracy for these classes. Most of the misclassifications in our cytogenetic domain were due to two sources. The first was the small numbers of D and R2 signals, which were inadequate for accurate estimation of densities for the relatively similar feature representations of the two classes. The second source of misclassification was between the D or R2 signals and the R1 signals due to the latter's numerical dominance in shaping the decision boundaries for the classes. In both cases, the high misclassification rate of signals of the minority classes undermined high overall accuracy for the imbalanced data.

We studied two solutions to diminish these problems and raise the classification accuracy. First was a hierarchical strategy that, in each hierarchy level, rendered simpler experts that were trained using classes of larger and

approximately equal sizes. This improved the accuracy and also provided insights into the difficulty in discriminating specific classes, thereby spotting some of the bottlenecks of the classification task.

The second solution we evaluated was balancing the data by up-sampling. We balanced the data for the NBC and MLP trained using the monolithic strategy and then assessed the usefulness of feature extraction in alleviating the problem of smallness of the data. Evaluating other classifiers to this problem (e.g., a support vector machine) is left to future study. Balancing by up-sampling the minority classes (R2 and D) improved the classification of patterns of these classes as the enlarged number of patterns enabled more accurate estimation of the densities. Note, however, that up-sampling minority classes for the MLP is equivalent to increasing the classifier training period, but only for these classes. This way, the MLP succeeds in improving its generalization capability without overfitting the entire data. Nevertheless, the main flaw of up-sampling when diminishing the imbalance problem is that the training and test sets share patterns in common and, thus, the accuracies of these sets cannot be considered independent anymore. Finally, it is interesting to see that the highest accuracy was achieved due to a combination of approaches, i.e., an accurate classifier (MLP) trained on a dimension-reduced balanced data set.

With respect to future research, we note that the moderate degree of improvement in accuracy that was achieved through balancing the data (Section 5) suggests that the smallness of the data is a more acute problem than the imbalance of the data. A similar conclusion was also made before [30]. Hence, one main goal of future research is to study these two problems separately. Another goal is the enlargement of the database with the emphasis on collecting patterns of minority classes as the impact of the skewed class prior probabilities on the overall accuracy is reduced with the data sample size [33]. In addition, we will use a panel of cytogenetic experts in order to minimize labeling errors of the signals used for training the classifiers. If up-sampling will still be needed, we would consider adding random noise to the up-sampled patterns in order to better represent the density using the finite sample size, although other alternatives for sampling may do as well [30]. We are also interested in evaluating the suggested solutions in other small, imbalanced domains in order to be able to comment more on the advantages and drawbacks of each solution. Finally, we plan to study the clinical implications of the proposed methodology for dot and non-dot-like FISH signal classification.

ACKNOWLEDGMENTS

This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University, Beer-Sheva, Israel.

REFERENCES

[1] S.J. Raudys and A.K. Jain, "Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, pp. 252-264, 1991.

[2] N. Japkowicz and S. Stephen, "The Class Imbalance Problem: A Systematic Study," *Intelligent Data Analysis,* vol. 6, pp. 429-450, 2002.

[3] M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection," *Proc. 14th Int'l Conf. Machine Learning,* pp. 179-186, 1997.

[4] D. Pinkel, J. Landegent, C. Collins, J. Fuscoe, R. Segraves, J. Lucas, and J. Gary, "Fluorescence in Situ Hybridization with Human Chromosome-Specific Libraries: Detection of Trisomy 21 and Translocations of Chromosome 4," *Proc. Nat'l Academy of Sciences,* vol. 85, pp. 9138-9142, 1988.

[5] J. Nath and K.L. Johnson, "A Review of Fluorescence in Situ Hybridization (FISH): Current Status and Future Prospects," *Biotech Histochemistry,* vol. 75, pp. 54-78, 2000.

[6] 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,* vol. 28, pp. 1-10, 1997.

[7] MetaSystems, Germany, http://www.metasystems.de, 2007.

[8] B. Lerner, W.F. Clocksin, S. Dhanjal, M.A. Hult'en, and C.M. Bishop, "Feature Representation and Signal Classification in Fluorescence In-Situ Hybridization Image Analysis," *IEEE Trans. Systems, Man, and Cybernetics, Part A,* vol. 31, pp. 655-665, 2001.

[9] B. Lerner, "Bayesian Fluorescence in Situ Hybridization Signal Classification," *Artificial Intelligence in Medicine,* vol. 30, pp. 301-316, 2004.

[10] B.R. Migeon, L.J. Shapiro, R.A. Norum, T. Mohandas, J. Axelman, and R.I. Dabora, "Differential Expression of Steroid Sulphatase Locus on Active and Inactive Human X Chromosome," *Nature,* vol. 299, pp. 838-840, 1982.

[11] T. Litmanovitch, M.M. Altaras, B. Dotan, and L. Avivi, "Asynchronous Replication of Homologous α-Satellite DNA Loci in Man Is Associated with Nondisjunction," *Cytogenetic Cell Genetics,* vol. 81, pp. 26-35, 1998.

[12] J. Yeshaya, R. Shalgi, M. Shoat, and L. Avivi, "FISH-Detected Delay in Replication Timing of Mutated FMR1 Alleles on Both Active and Inactive X-Chromosomes," *Human Genetics,* vol. 105, pp. 86-97, 1999.

[13] B. Lerner, S. Dhanjal, and M.A. Hult'en, "GELFISH—Graphical Environment for Labelling FISH Images," *J. Microscopy,* vol. 203, pp. 258-268, 2001.

[14] J.P. Serra, *Image Analysis and Mathematical Morphology.* Academic Press, 1982.

[15] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 9, pp. 62-66, 1979.

[16] L. Vincent and P. Soille, "Watershed in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, pp. 583-598, 1991.

[17] C.O. de Solorzano, A. Santos, I. Vallcorba, J.M. Garcia-Sagredo, and F. del Pozo, "Automated FISH Spot Counting in Interphase Nuclei: Statistical Validation and Data Correction," *Cytometry,* vol. 31, pp. 93-99, 1998.

[18] K.R. Castleman, *Digital Image Processing.* Prentice-Hall, 1996.

[19] P. Soille, *Morphological Image Analysis: Principles and Applications.* Springer-Verlag, 1999.

[20] E.S. Baugher and A. Rosenfeld, "Boundary Localization in an Image Pyramid," *Pattern Recognition,* vol. 19, pp. 373-395, 1986.

[21] A. Likas, N. Vlassis, and J.J. Verbeek, "The Global *k*-Means Clustering Algorithm," *Pattern Recognition,* vol. 36, pp. 451-461, 2003.

[22] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification.* Wiley, 2001.

[23] J. Iivarinen, M. Peura, J. Sarela, and A. Visa, "Comparison of Combined Shape Descriptors for Irregular Objects," *Proc. Eighth British Machine Vision Conf. (BMVC '97),* vol. 2, pp. 430-439, 1997.

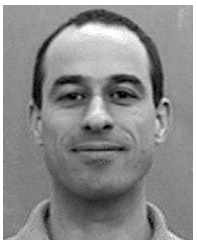
[24] P.A. Devijver and J. Kittler, *Pattern Recognition—A Statistical Approach.* Prentice Hall, 1982.

[25] P. Langley, W. Iba, and K. Thompson, "An Analysis of Bayesian Classifiers," *Proc. 10th Nat'l US Conf. Artificial Intelligence,* pp. 223-228, 1992.

[26] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," *Machine Learning,* vol. 29, pp. 131-163, 1997.

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

[28] B.W. Silverman, *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, 1986.
[29] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc. B,* vol. 39, pp. 1-38, 1977.
[30] G.E.A.P.A. Batista, R.C. Prati, and M.C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *SIGKDD Explorations,* vol. 6, pp. 20-29, 2004.
[31] T.G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation,* vol. 10, pp. 1895-1924, 1998.
[32] E. Alpaydin, "Combined 5x2cv *F* Test for Comparing Supervised Classification Learning Algorithms," *Neural Computation,* vol. 11, pp. 1885-1892, 1999.
[33] G.M. Weiss and F. Provost, "Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction," *J. AI Research,* vol. 19, pp. 315-354, 2003.



Boaz Lerner received the BA degree in physics and mathematics from the Hebrew University, Israel, in 1982 and the PhD degree in computer engineering from Ben-Gurion University, Israel, in 1996. He performed research at the Neural Computing Research Group at Aston University, Birmingham, United Kingdom, and the Computer Laboratory of the University of Cambridge, Cambridge, United Kingdom. In 2000, he joined the Department of Electrical and Computer Engineering at Ben-Gurion University, Israel, where he is currently a senior lecturer. His current interests include machine learning approaches to data analysis, learning Bayesian networks, neural networks, and their application to "real-world" problems.

Josepha Yeshaya received the BA degree in chemistry from Tel-Aviv University, Israel, in 1985 and the PhD degree in human genetics from the same university in 1999. During 1992 to 1993, she performed research at the Molecular Genetic Laboratory of the Department of Human Genetics, Tel-Hashomer Hospital, Israel. Since 1999, she has been the head of the Cytogenetic Laboratory at the Rabin Medical Center of the Beilinson Campus, Petah-Tiqva, Israel, where she also conducts research. Her fields of interest include the detection of genomic imbalances by replication timing of genes and mode of behavior of genes.

Lev Koushnir received the BSc degree from the Technion, Israel, in 1999. He is currently pursuing the MSc degree at Ben-Gurion University, Israel.