The Bayesian ARTMAP

Boaz Vigdor and Boaz Lerner, *Member, IEEE*

***Abstract*—In this paper, we modify the fuzzy ARTMAP (FA) neural network (NN) using the Bayesian framework in order to improve its classification accuracy while simultaneously reduce its category proliferation. The proposed algorithm, called Bayesian ARTMAP (BA), preserves the FA advantages and also enhances its performance by the following: 1) representing a category using a multidimensional Gaussian distribution, 2) allowing a category to grow or shrink, 3) limiting a category hypervolume, 4) using Bayes' decision theory for learning and inference, and 5) employing the probabilistic association between every category and a class in order to predict the class. In addition, the BA estimates the class posterior probability and thereby enables the introduction of loss and classification according to the minimum expected loss. Based on these characteristics and using synthetic and 20 real-world databases, we show that the BA outperformes the FA, either trained for one epoch or until completion, with respect to classification accuracy, sensitivity to statistical overlapping, learning curves, expected loss, and category proliferation.**

***Index Terms*—Bayes' decision theory, category proliferation, classification, fuzzy ARTMAP (FA), neural network (NN).**

I. Introduction

THE fuzzy ARTMAP (FA) is considered as one of the leading neural network (NN) algorithms for classification [1]. The FA excels in fast incremental supervised learning in a nonstationary environment. The network allows learning new data without forgetting past data, tackling the so-called "plasticity–stability dilemma" [2], which is crucial in incremental learning. Following increase in data complexity, the FA expands its complexity by allocating nodes dynamically without user intervention. In addition, the algorithm depends on minimal heuristics and settings of parameters and guarantees short training periods and convergence [1], [3], [4]. The FA and its variants have been found accurate and fast learners as exemplified in performing various classification tasks, such as automatic target recognition based on radar range profiles [5], speaker-independent vowel recognition [6], online handwritten recognition [7], electrocardiogram (ECG) signal recognition [8], medical diagnosis of breast cancer and heart disease [9], 3-D object understanding and prediction from a series of 2-D views [10], and recently also genetic abnormality diagnosis [11].

The major drawback of the FA is its sensitivity to statistical overlapping between the classes. It is the self-organization nature of the algorithm that while enabling continuous learning of novel patterns also overfits noisy (overlapped) data that are mistakenly considered as novel. This sensitivity is responsible to uncontrolled growth in the number of categories, also referred to as category proliferation, leading to high computational and memory complexities and possible degradation in the classification accuracy [4], [11], [12]. In order to tackle category proliferation and to improve other characteristics of the FA, researchers have proposed several enhancements to the FA such as PROBART [4], Gaussian ARTMAP (GA) [6], ARTMAP-IC [9], and ART-EMAP [10], as well as modifications to the FA methodology [12], [13]. The improvement to classification accuracy and reduction in category proliferation due to these modifications were examined on several synthetic and real-world databases [6], [12], [13].

We propose the Bayesian ARTMAP (BA) that modifies some of the characteristics of the FA algorithm by the following: 1) replacing the hyperrectangular category with a Gaussian category, 2) limiting the volume of a selected category hence allowing the category to grow or shrink, 3) associating patterns with categories and categories with classes probabilistically in order to perform, respectively, ART and ARTMAP learning, and 4) enabling class probabilistic inference using all the associated categories. The BA also estimates category and class posterior probabilities and thus enables the introduction of different losses into the classification task. The BA is evaluated here in comparison to the FA using different experiments and synthetic and real-world data. The classifiers are evaluated with respect to their classification accuracy, rise in the number of categories, learning curves, expected loss, and sensitivity to statistical overlapping. For all these criteria and in all the experiments, the BA proves superior performance to the FA.

Section II briefly summarizes the principles and dynamics of the FA, whereas Section III introduces Bayesian ART and BA learning and inference. Section IV extensively compares the characteristics of the BA, FA, and GA [6]. In Section V, these classifiers as well as others are experimentally and thoroughly compared before completing the paper in Section VI with a discussion.

II. FA Principles and Dynamics

The FA NN for incremental supervised learning [1] incorporates two fuzzy adaptive resonance theory (ART) [2] modules denoted as ART_a and ART_b that are linked by a map field module associating nodes (categories) from ART_a with nodes in ART_b. The fuzzy ART module [14] performs fast incremental unsupervised learning by clustering M-dimensional input patterns (every input pattern $\boldsymbol{I} = (I_1, I_2, \ldots, I_3)$ is initially complement-coded [1]) into categories, each forming a hyperrectangular region in the M-dimensional input space. The jth category is defined by a vector of weights w_j that each of its elements

Manuscript received November 14, 2005; revised October 19, 2006; accepted March 1, 2007. This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University, Beer-Sheva, Israel.

The authors are with the Pattern Analysis and Machine Learning Laboratory, Department of Electrical Computer Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel (e-mail: boaz@ee.bgu.ac.il).

Digital Object Identifier 10.1109/TNN.2007.900234

is initially set at 1 and is monotonically nonincreasing through time.

Categorization with the fuzzy ART is performed in three stages: category choice, category match (vigilance test), and learning. In the category choice stage, a choice function is calculated for the current pattern $\boldsymbol{I}$ and each existing category

$$T_j = \frac{||\boldsymbol{I} \wedge \boldsymbol{w}_j||}{\alpha + ||\boldsymbol{w}_j||} \tag{1}$$

where $\wedge$ is the fuzzy AND operation $(\mathbf{X} \wedge \mathbf{Y})_i = \min(x_i, y_i)$, $\alpha > 0$ is a choice parameter[1] and the norm is L_1. The chosen category is the one achieving the highest value of the choice function.

When a category J is chosen, a hypothesis test called the vigilance test is performed in order to measure the category match to the pattern. If the match function exceeds the vigilance parameter $\rho \in [0, 1]$

$$\frac{||\boldsymbol{I} \wedge \boldsymbol{w}_J||}{||\boldsymbol{I}||} \geq \rho \tag{2}$$

then the chosen category is said to win (match) and learning is performed. Otherwise, the chosen category is removed from the search by forcing the value of T_J to zero for the rest of this pattern presentation. As a result, a new category maximizing the choice function (1) is chosen and the process continues until a chosen category satisfies the vigilance test (2). If none of the existing categories meets the vigilance test, a new category is formed and learning for this category is performed without a vigilance test. Either way, learning in the fuzzy ART is accomplished by updating the weight vector of the winning (or new) category according to

$$\boldsymbol{w}_J^{\text{new}} = \beta \left(\boldsymbol{I} \wedge \boldsymbol{w}_J^{\text{old}}\right) + (1 - \beta)\boldsymbol{w}_J^{\text{old}} \tag{3}$$

where $\beta \in (0, 1]$ is the learning rate and $\beta = 1$ defines fast learning. Note that the vigilance parameter controls the similarity required between the chosen category and the input pattern (2) in order to control learning. Lowering the vigilance parameter provides broader generalization (large categories) and vice versa.

Considering the FA for pattern recognition, the input to ART_a is the pattern to classify and the input to ART_b is the pattern label. The map field includes a matrix of weights w^{ab} which maps ART_a categories to ART_b categories. The Jth row vector of w^{ab} denotes the prediction of ART_b categories as a result of the Jth winning category in ART_a. During the training phase, the map field performs a vigilance test similarly to that of ART_a, where if the match function exceeds the map field vigilance parameter $\rho_{ab} \in [0, 1]$, then learning occurs. This test assures that the prediction of the correct class complies with the label represented by the winning ART_b category. Else, a match tracking procedure is activated for finding a better category in ART_a. In this process, the map field raises ART_a vigilance parameter ρ_a until the current Jth category fails ART_a vigilance test (2) and is removed from the competition. The search in ART_a proceeds until an ART_a category that predicts the correct ART_b category is chosen; otherwise, a new category is created. When the ART_a Jth category upholds the map vigilance test, its association to the ART_b categories is adapted by the following learning rule:

$$\boldsymbol{w}_J^{ab,\text{new}} = \beta_{ab}\left(\boldsymbol{w}_J^{ab,\text{old}} \wedge \boldsymbol{y}^b\right) + (1 - \beta_{ab})\boldsymbol{w}_J^{ab,\text{old}} \tag{4}$$

where the components of ART_b output y^b are zero except the Kth component which is 1 if the Kth category wins in ART_b.

In fast learning mode ($\beta_{ab} = 1$), the link between the ART_a Jth category and the ART_b Kth category becomes permanent, i.e., $w_{JK}^{ab} = 1$ for all data presentations. In the test phase, only ART_a is active so the vigilance test in the map field is avoided. The class prediction is deduced from the map field weights of the winning ART_a category.

III. BA Algorithm

The BA is an alternative to the FA utilizing the latter fast incremental learning, however, also diminishing its main shortcoming which is category proliferation. Category proliferation in the FA originates mainly from two sources—inadequate representation of the data and sensitivity to class overlapping. The inadequacy of the representation is derived from using the fuzzy set theory "minimum $(\wedge)$" and "maximum $(\vee)$" operators that lead to categories having hyperrectangular class decision boundaries. A hyperrectangle may suit data distributed uniformly but not the most natural (Gaussian) data distribution that solicits a decision boundary in the form of a hypersphere or hyperellipsoid. As the dimension of the classification problem increases, the ratio between the volumes of a hyperrectangle and hyperellipsoid (both bounding the data) increases monotonically [6]. That is, as the dimension increases, the hyperrectangle category represents higher volumes in which there are no patterns to support this representation (these are the hyperrectangle "corners"). Furthermore, patterns residing in these corners but belonging to different classes of the class that is associated with the category, are wrongly clustered by the category. In such a case, a match tracking procedure is activated for finding a better existing category and if not found, a new category is created. As class overlapping increases, the number of such activations increases leading to category proliferation.

The second source of category proliferation is the FA sensitivity to statistical overlapping between classes. Overlapping is responsible for misclassifications during FA training. Each misclassification requires match tracking and raising the vigilance parameter in order to find a more suitable category for the misclassified pattern. The selected category needs a larger weight vector in order to beat the new vigilance parameter in the vigilance test (2), and thus also a smaller size. Repeated for many patterns of both classes and because small categories cannot represent wide regions, these misclassifications are responsible for a large number of small categories within the overlapping area. Moreover, if no existing category is found for a misclassified pattern during match tracking, a new category is formed. Either way, the result is category proliferation that is intensified with the degree of class overlapping.

The proposed BA employs the main stages of the FA but it replaces the deterministic rules of the FA with statistical learning and inference. There are several main differences between the

[1]An elaborated examination of the role of the choice parameter can be found in [15].

FA and BA. First is the shape of the categories—hyperrectangle for the FA versus multidimensional Gaussian for the BA. Generally, the use of multidimensional Gaussian categories provides better representation of the naturally distributed data. It also requires a fewer categories and provides enhanced flexibility in the representation [6]. Second is the category choice function (1) that is based on fuzzy set-theory operations (FA) versus Bayes' decision theory (BA). Bayes' decision theory accounts not only to the distance of a category to a pattern but also to the dominance of the category with respect to other categories through the category prior probability. The third difference is the match function (2) limiting the size (FA) versus volume (BA) of a category. Characterization using the volume (product of sides[2]) instead of size (sum of sides) is more appropriate for high-dimensional clusters in real-world domains [16]. Fourth is class prediction based on a single category (FA) versus many categories (BA). Basing class prediction on all the categories that are probabilistically associated with the class rather than on the single winning category provides better generalization. These main differences between the FA and BA, as well as others, will be elaborated in Section IV, where we will also extend on the similarities and differences between the BA and GA [6].

The BA is composed of hierarchies, as the FA. The clustering algorithm, called Bayesian ART, is described in Section III-A whereas the BA in Section III-B.

A. Bayesian ART

Assuming the existence of a (conditional) probability density function for each class, we replace the deterministic fuzzy ART hyperrectangular category by a multidimensional Gaussian component and represent the class density using a mixture of such components. Each Gaussian category is fully characterized by its mean vector, covariance matrix, and prior probability. These parameters provide extended information about the category compared to the weight vector of the fuzzy ART hyperrectangular category. That is, instead of a vague idea about a category as exemplified for the fuzzy ART by a weight vector composing of the category two extreme corners [1], the Gaussian category of the BA is clearly identified by its central of mass, shape of distribution, and dominance with respect to other clusters. For example, a category of the fuzzy ART may have the same parameters (weights) whether it clusters two patterns or two thousand patterns, whereas using the Bayesian ART the category would have different prior probabilities. Furthermore, the distribution of the data within the category in the fuzzy ART is completely unknown, where for the BA this distribution is estimated using a Gaussian represented by the data covariance matrix. Also, with Gaussian representation there is no need in complement coding as for the FA (Section II).

Similarly to the fuzzy ART, the Bayesian ART algorithm composes of three main stages, namely, category choice, category match (vigilance test), and learning.

1) Category Choice: In this stage, all existing categories compete to represent an input pattern. The *a posteriori* probability of the jth category to represent the M-dimensional pattern $\boldsymbol{x}$[3] is computed by

$$M_j = \hat{P}(w_j|\boldsymbol{x}) = \frac{\hat{p}(\boldsymbol{x}|w_j)\hat{P}(w_j)}{\sum_{l=1}^{N_{\text{cat}}} \hat{p}(\boldsymbol{x}|w_l)\hat{P}(w_l)} \tag{5}$$

where N_{cat} is the number of categories and $\hat{P}(w_j)$ is the estimated prior probability of the jth category. The likelihood of w_j with respect to $\boldsymbol{x}$ is estimated using all patterns that have already been associated with the multivariate Gaussian category w_j

$$\hat{p}(\boldsymbol{x}|w_j) = \frac{1}{(2\pi)^{M/2}\left|\hat{\Sigma}_j\right|^{1/2}} \times \exp\left\{-0.5\left(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j\right)^T \hat{\Sigma}_j^{-1}\left(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j\right)\right\} \tag{6}$$

where $\hat{\boldsymbol{\mu}}_j$ and $\hat{\Sigma}_j$ are the estimated mean and covariance matrix of the jth category.

The chosen (i.e., winning) category J is the one with the maximum *a posteriori* probability (MAP)

$$J = \arg\max_j(M_j). \tag{7}$$

That is, the Jth category w_J is either more populated than other categories [i.e., having high $\hat{P}(w_j)$] or more likely to be the true category for $\boldsymbol{x}$ [i.e., having high $\hat{p}(\boldsymbol{x}|w_j)$ since, e.g., it is the closest category to $\boldsymbol{x}$] or both. Based on both probabilities and Bayes' theorem [16], the MAP criterion is expected to select a winning category for the BA more accurately than if using only one of the probabilities. For example, the MAP criterion may prefer a category having *a priori* probability which is higher than that of another category although the normalized by the covariance distance[4] of the former to the pattern is larger than that of the latter.

2) Category Match (Vigilance Test): Similarly to the FA, the purpose of the vigilance test is to ensure that the chosen category is limited in size. The test restricts the BA Jth category hypervolume[5] S_J to the maximal hypervolume allowed for a category S_{MAX}

$$S_J \le S_{\text{MAX}} \tag{8}$$

where the hypervolume is defined as the determinant of the Gaussian covariance matrix. For a diagonal covariance matrix, this hypervolume is reduced to the product of the variances each for a dimension

$$S_J \triangleq \det(\Sigma_J) = \prod_{d=1}^{M} \sigma_{J_d}^2. \tag{9}$$

[2]In the BA, the category side along a dimension is represented by the standard deviation of the Gaussian along this dimension.

[3]In Section II, we denoted a pattern by $\boldsymbol{I}$ to comply with the FA literature [1]. We also denote the jth category by w_j.

[4]This is the argument of the exponent in (6), which is the Mahalanobis distance [16].

[5]For multidimensional categories, the volume is the natural way to represent category size [16].

If the winning category matches this criterion (8), learning is performed. Else, the category is removed from the competition for this pattern and the Bayesian ART searches for another category until finding one complying with (8). If all existing categories fail the vigilance test, a new category is formed having a mean vector which is the input pattern and an initial covariance matrix Σ_{init} that enables meeting (8) (it will be elaborated in Experiment 1 in Section V-A).

3) Learning: When a chosen category matches the maximal hypervolume (8), then the category parameters are adjusted by the following update equations:

$$\hat{\boldsymbol{\mu}}_{J,\text{new}} = \frac{N_J}{N_J+1}\hat{\boldsymbol{\mu}}_{J,\text{old}} + \frac{1}{N_J+1}\boldsymbol{x} \tag{10}$$

$$\hat{\Sigma}_{J,\text{new}} = \frac{N_J}{N_J+1}\hat{\Sigma}_{J,\text{old}} + \frac{1}{N_J+1}\left(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_{J,\text{new}}\right) \times \left(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_{J,\text{new}}\right)^T * I \tag{11}$$

where N_J is the number of patterns that have been clustered by the Jth category before introducing the current pattern and I is the identity matrix. These mean and covariance matrix update equations are expanded to the multidimensional case from sequential maximum–likelihood estimation for a single Gaussian. The element product $*$ in (11) is performed if a diagonal matrix is assumed. Updating the category prior probability is discussed in Section III-B1.

B. Bayesian ARTMAP

1) Learning: In the training phase, the BA performs two functions similarly to the FA but with some changes. First, the BA maps each category of the Bayesian ART to a class. This mapping can be deterministic like in the fast learning mode of the FA, i.e., each category is mapped to only one class, or it can be probabilistic, i.e., category w_j is assigned to class c_i with probability $P(c_i|w_j)$. $P(c_i|w_j)$ is the probabilistic alternative to the deterministic matrix of weights w^{ab} of the FA.

The second similar function is match tracking but using a different criterion. As recalled, the map field in the FA compares the prediction of the class to the pattern label. If the match is below a certain threshold, then mismatch occurs and a new category is searched for after raising ρ_a. In fast learning, this means that the FA should predict each training pattern correctly. This criterion is highly sensitive to statistical overlapping between different classes and it produces a large number of small categories in the overlapping region. In the BA, the proposed criterion utilizes the class posterior probability for the winning category $P(c_i|w_J)$. If this probability is larger than a threshold $P_{\min}$ (in analogy to the FA)

$$P(c_i|w_J) \geq P_{\min} \tag{12}$$

then the Jth category is associated with the ith class. Otherwise, mismatch occurs and match tracking (similar to that performed by the FA) is activated by lowering the category maximal hypervolume S_{MAX}

$$S_{\text{MAX}} = S_J - \delta \quad 0 < \delta \ll S_J \tag{13}$$

enough to disqualify the winning Jth Bayesian ART category. As a result, the Bayesian ART starts searching for a new category that must have a smaller hypervolume S_J. This procedure continues until a winning category satisfies (12) or a new category is formed.

Learning by the BA is carried out by estimating the joint category-and-class probability $\hat{P}(c_i, w_j)$ using the frequency count. The count is stored in a matrix $[N_{ij}]_{C*N_{\text{cat}}}$ holding in the ijth entry the number of training patterns that belong to the ith class of the C classes and are clustered to the jth category of the N_{cat} categories.[6] $\hat{P}(c_i, w_j)$ is estimated by

$$\hat{P}(c_i, w_j) = \frac{N_{ij}}{\sum_{k=1}^{C}\sum_{l=1}^{N\text{cat}} N_{kl}}. \tag{14}$$

Marginalizing this joint probability over the classes, we get the jth category prior probability that is used in the computation of the *a posteriori* probability of the jth category (5)

$$\hat{P}(w_j) = \sum_{k=1}^{C}\hat{P}(c_k, w_j) = \frac{\sum_{k=1}^{C} N_{kj}}{\sum_{k=1}^{C}\sum_{l=1}^{N\text{cat}} N_{kl}}. \tag{15}$$

Using Bayes' theorem, the estimate for the class posterior probability given a category is

$$\hat{P}(c_i|w_j) = \frac{\hat{P}(c_i, w_j)}{\hat{P}(w_j)} = \frac{\frac{N_{ij}}{\sum_{k=1}^{C}\sum_{l=1}^{N\text{cat}} N_{kl}}}{\frac{\sum_{k=1}^{C} N_{kj}}{\sum_{k=1}^{C}\sum_{l=1}^{N\text{cat}} N_{kl}}} = \frac{N_{ij}}{\sum_{k=1}^{C} N_{kj}} \tag{16}$$

which is simply the number of patterns from the jth category that are associated to the ith class normalized by the number of patterns clustered by the jth category. Finally, when an input pattern belonging to class c is learned by the winning Jth Bayesian ART category, the frequency count for the corresponding matrix entry is updated

$$N_{cJ}^{\text{new}} = N_{cJ}^{\text{old}} + 1. \tag{17}$$

2) Inference: Inference corresponds to the association of a category to a class when predicting a test pattern. During the test, the FA declares the winning class as the class associated with the winning ART_a category for a test pattern. In contrast, the BA performs inference by using all the categories that are associated to the class. That is, the class chosen for a test pattern $\boldsymbol{x}$ is

$$c_I = \arg\max_i \hat{P}(c_i|\boldsymbol{x}) \tag{18}$$

[6] In contrast, categories and classes in the FA are represented, respectively, by the rows and columns of w^{ab}.

where

$$\begin{aligned}\hat{P}(c_i|\boldsymbol{x}) &= \frac{\hat{P}(c_i,\boldsymbol{x})}{\hat{p}(\boldsymbol{x})} = \frac{\sum_{j=1}^{N\text{cat}} \hat{P}(c_i,w_j,\boldsymbol{x})}{\hat{p}(\boldsymbol{x})} \\ &= \frac{\sum_{j=1}^{N\text{cat}} \hat{P}(c_i|w_j,\boldsymbol{x})\hat{P}(w_j,\boldsymbol{x})}{\hat{p}(\boldsymbol{x})} \\ &= \frac{\sum_{j=1}^{N\text{cat}} \hat{P}(c_i|w_j)\hat{p}(\boldsymbol{x}|w_j)\hat{P}(w_j)}{\sum_{k=1}^{C}\sum_{l=1}^{N\text{cat}} \hat{P}(c_k|w_l)\hat{p}(\boldsymbol{x}|w_l)\hat{P}(w_l)} \end{aligned} \tag{19}$$

where $\hat{P}(c_i|w_j)$, $\hat{P}(w_j)$, and $\hat{p}(\boldsymbol{x}|w_j)$ have already been defined in (16), (15), and (6), respectively. In the fourth equality of (19), we assume that

$$\hat{P}(c_i|w_j,\boldsymbol{x}) = \hat{P}(c_i|w_j) \tag{20}$$

which means that once the Bayesian ART identifies the winning category for the test pattern, it is only the association between the category and class that affects the classification of the pattern. This assumption is also made by the FA.

Take, for example, the following hypothetical matrix derived at the end of training:

$$[N_{ij}]_{C*N_{\text{cat}}} = \begin{bmatrix} 1 & 0 & 7 & 12 & 0 \\ 0 & 5 & 0 & 12 & 0 \\ 3 & 0 & 0 & 0 & 20 \end{bmatrix}.$$

It represents a frequency count for three classes and a BA having five categories. The estimated probabilities based on (14)–(16) are, respectively

$$\hat{P}(c_i,w_j) = \frac{1}{60}\begin{bmatrix} 1 & 0 & 7 & 12 & 0 \\ 0 & 5 & 0 & 12 & 0 \\ 3 & 0 & 0 & 0 & 20 \end{bmatrix}$$

$$\hat{P}(w_j) = \frac{1}{60}\begin{bmatrix} 4 & 5 & 7 & 24 & 20 \end{bmatrix}$$

$$\hat{P}(c_i|w_j) = \begin{bmatrix} \frac{1}{4} & 0 & 1 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & \frac{1}{2} & 0 \\ \frac{3}{4} & 0 & 0 & 0 & 1 \end{bmatrix}.$$

As this example shows, a category may represent more than one class but the probabilities of doing so should summed to 1, as required from posterior probabilities. Also, the example shows that categories 2, 3, and 5 are associated entirely with classes 2, 1, and 3, respectively. Category 1 is mainly associated with class 3 (having a posterior of 3/4) and slightly with class 1 (having a posterior of 1/4), whereas category 4 is equally associated with classes 1 and 2. These probabilities reflect the contribution of each category to the selection of each class when computing (18). Note that in almost half of the cases in this example there is no association between a class and a category.

IV. Related Methodologies

The BA is closely related to the FA [1] and GA [6]. The PROBART algorithm, although having probabilistic characteristics, has its strength in data projection and not in classification [4] and, therefore, is not considered here. Both the BA and GA are based on the FA, where each of the models go through the same stages, namely category choice, category match, and learning for the ART module and training and inference (test) for the ARTMAP module. The differences are in the implementation of these stages, where generally both the GA and BA are dissimilar from the FA. Although the BA and GA have been developed independently, they share several characteristics but at the same time also differ in some other characteristics. Table I compares the main characteristics of the three models.

The main differences between the FA and the GA/BA models are, respectively, as follows: 1) pattern normalization through complement coding versus Mahalanobis distance (the distance from the pattern to the category mean normalized by the category variance); 2) hyperrectangular versus multidimensional Gaussian categories; also, category determination is in a "hard" fashion as the smallest hyperrectangle that encloses all category patterns versus a "soft" fashion as determined by the natural decline of the multidimensional Gaussian; 3) categories can only grow versus grow or shrink as a consequence of statistical learning; 4) category choice using terms of fuzzy set theory operations versus Bayes' decision theory; using Bayes' theory both the GA and BA favors categories that are either close to the pattern and small (through the likelihood), highly populated (through the prior probability), or both; 5) ART learning by category movement towards the pattern in terms of fuzzy set theory operations versus terms of maximum–likelihood-based sequential updating of parameters (mean, covariance, and prior probability); and 6) class prediction during inference is based on a single versus multiple category(ies) associated with the class.

The differences between the GA and BA are in both the ART and ARTMAP modules. Following Table I, we identify the first difference in the multidimensional representation of the Gaussian cluster. It is limited in the GA by a diagonal covariance matrix; however, it is not limited in the BA allowing any covariance matrix. That is, the first model implicitly assumes feature independence, whereas the second model does not restrict feature description.

The second difference is in the choice function. The GA employs a discriminant function (i.e., the logarithm of the joint probability), whereas the BA computes the posterior probability and thus establishes a generative model. When we are only interested in the classification accuracy, the two models provide the same; however, the BA model is more flexible and general due to the computation of the posterior probability. Often, it is desirable to introduce different losses to misclassifications or to have the possibility of rejecting a pattern without classifying it or to correct different class prior probabilities in the training and test sets [17]. Using posterior probabilities, these are natural to the BA but infeasible when employing the GA.

The third difference between the models affects directly the performance of the models. During category match, the GA determines how well the winning category matches the pattern

TABLE I
MAIN CHARACTERISTICS OF THE FA, GA, AND BA CLASSIFIERS

	FA [1]	GA [6]	BA
Initial parameters	choice $\boldsymbol{\alpha > 0}$, vigilance $\boldsymbol{0 \le \rho, \rho_{ab} < 1}$	initial standard deviation (γ), $\boldsymbol{\rho = 0}$	initial maximal hypervolume ($\boldsymbol{S_{MAX}}$), $\boldsymbol{P_{min}}$
Categories	hyperrectangle	multidimensional Gaussian (diagonal covariance matrix)	multidimensional Gaussian (any covariance matrix)
Category size	can only grow	can grow or shrink	can grow or shrink
Training modes	on-line, with validation and until completion	on-line	on-line
Choice function	$\boldsymbol{T_j = \frac{\|\|I \wedge w_j\|\|}{\alpha + \|\|w_j\|\|}}$	$\boldsymbol{\log((2\pi)^{M/2}\hat{p}(x\|w_j)\hat{P}(w_j))}$	$\boldsymbol{\hat{P}(w_j\|x) = \frac{\hat{p}(x\|w_j)\hat{P}(w_j)}{\sum_{l=1}^{N_{cat}} \hat{p}(x\|w_l)\hat{P}(w_l)}}$
Category match	$\boldsymbol{\frac{\|\|I \wedge w_J\|\|}{\|\|I\|\|} \ge \rho}$	$\boldsymbol{-(1/2)\sum_{d=1}^{M}(\frac{\mu_{Ji}-x_i}{\sigma_{Ji}})^2 \ge \rho}$	$\boldsymbol{S_J = \prod_{d=1}^{M}\sigma_{J_d}^2 \le S_{MAX}}$
ART learning	$\boldsymbol{w_J^{new} = I \wedge w_J^{old}}$ (for $\boldsymbol{\beta = 1}$)	as for the BA but for diagonal covariance matrices	$\boldsymbol{\hat{\mu}_{J,new} = \frac{N_J}{N_J+1}\hat{\mu}_{J,old} + \frac{1}{N_J+1}x}$ $\boldsymbol{\hat{\Sigma}_{J,new} = \frac{N_J}{N_J+1}\hat{\Sigma}_{J,old} + \frac{1}{N_J+1}(x-\hat{\mu}_{J,new})(x-\hat{\mu}_{J,new})^T * I}$ $\boldsymbol{\hat{P}(w_j) = \sum_{k=1}^{C} N_{kj} / \sum_{k=1}^{C}\sum_{l=1}^{N_{cat}} N_{kl}}$
ARTMAP learning	$\boldsymbol{w_J^{ab,new} = w_J^{ab,old} \wedge y^b}$ (for $\boldsymbol{\beta_{ab} = 1}$)	mapping the winning category to the pattern class, unless this category have already been assigned to another class	mapping all categories to all classes in probability $\boldsymbol{\hat{P}(c_i\|w_j)}$
Inference	winning category	choosing the class that maximizes the sum of joint probabilities for each category associated with the class	as for the GA but each joint probability is weighted by $\boldsymbol{\hat{P}(c_i\|w_j)}$

using minus the squared Mahalanobis distance from the pattern to the mean of the category (denoted here as $-r^2$).[7] If $-r^2$ exceeds the vigilance parameter (Table I), the category resonates, i.e., the winning category matches the pattern and learning occurs by changing the category parameters in order to absorb the pattern into the category. Employing $-r^2$ emphasizes the importance that is given to the closeness of the winning category to the input pattern in evaluating their match. However, this category may have reached resonance only because it is large (having high standard deviation). That is, a large category (that its mean is not necessarily close to the pattern) has higher chances of meeting the category match.[8] Since a large category is more likely to represent patterns of different classes (especially in regions of class overlapping), this category will quite often cause mismatch during ARTMAP training followed by match tracking in order to find or establish a more suitable category. The consequence is category proliferation. Category proliferation could have been controlled if limiting the size of a category was part of the category match stage. This is indeed implemented by the BA. Only a winning category that its volume[9] is limited can meet the category match and be learned. Thus, limiting category proliferation is vital in order to stabilize FA-based models.

The fourth difference between the GA and BA is in the ARTMAP learning. Learning in the GA is accomplished by mapping the winning category to the pattern class unless this category has already been associated with another class. Learning in the BA is achieved by mapping all categories to all classes in probability (although most of the probabilities are zero, as is evident in the example of Section III-B2). Two or more categories in the BA may be associated with the same class (as in the GA) but also a single category may be associated (probabilistically) with more than a single class. Moreover, by associating all categories to all classes probabilistically, the BA becomes less sensitive than the GA to the order of presentation of training patterns. The GA depends on this order since a category that represents a pattern and associated with the pattern class could have been associated with another class if a pattern

[7]As is shown in (6) and Table I, the Mahalanobis distance [16] is proportional to the squared distance between the pattern and mean of the category and inversely proportional to the category size manifested by the standard deviations (or determinant of the covariance matrix).

[8]This category should also have high prior probability in order to win first the category choice stage.

[9]The BA defines the category volume by the determinant of the covariance matrix or, for a diagonal matrix, by the "product of category sides" (9), whereas the GA implicitly considers category size using $-r^2$, which is a sort of "sum of category sides."

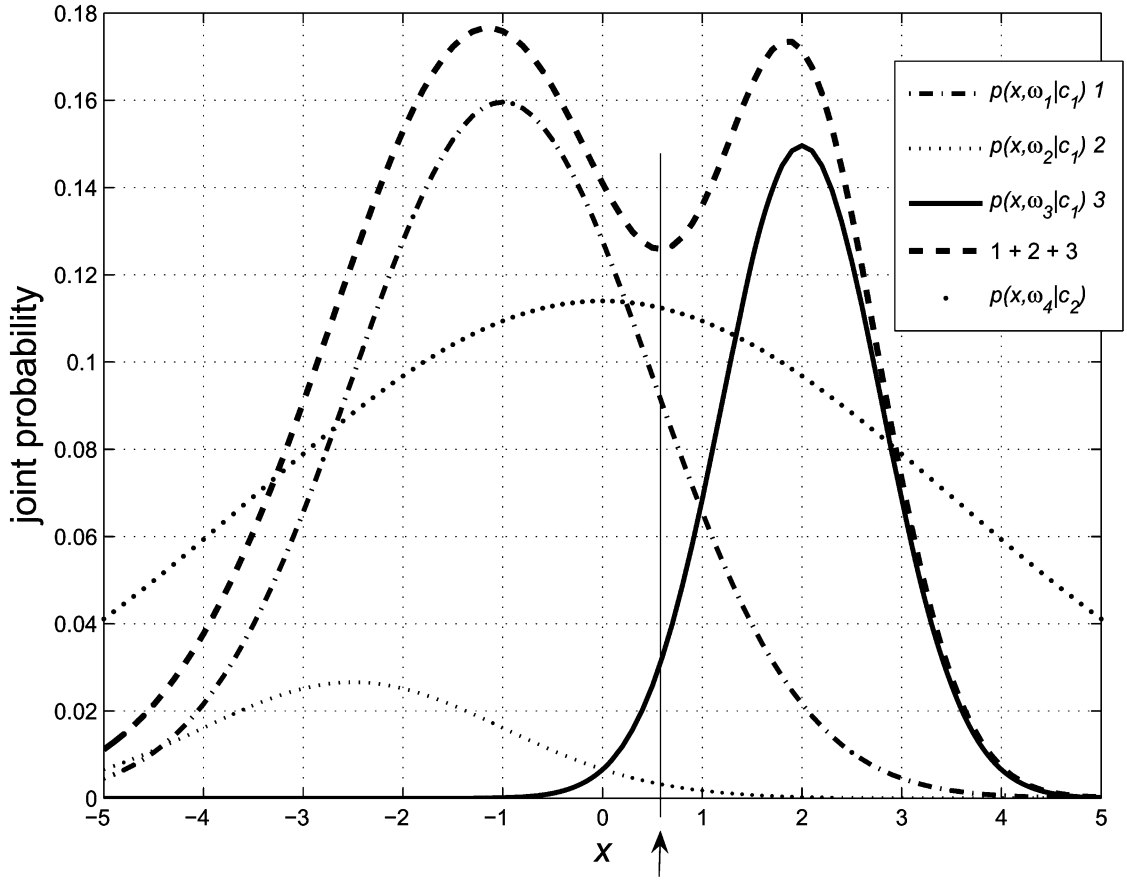

Fig. 1. Decision on a class based on the sum of unweighted pattern-category joint probabilities. A pattern indicated by the arrow will be classified to c_1 only because the unweighted sum of joint probabilities for this class is higher than the single joint probability for class c_2. Note, however, that the latter joint probability is the highest of all probabilities.

close to the former but from another class would have been presented first. This is especially significant in domains having significant class overlapping.

Last is the difference in the inference stage. The GA assigns a pattern to the class having the highest sum of joint probabilities for the pattern and each category which is associated with the class. The BA does the same but using the weighted and normalized sum. Normalization turns the sum of joint probabilities into posterior probability $P(c_i|x)$(19) and thereby enables the calculation of the probability that the test pattern indeed belongs to a class rather than just deciding that it belongs to this class. Weighting the joint probabilities by $\hat{P}(c_i|w_j)$ before the summation in (19) ensures that categories that are strongly associated with the class (as estimated during training) and thereby leading to higher posterior probabilities $\hat{P}(c_i|w_j)$ will influence the selection of the class more than other categories that are marginal to the class. That is, categories representing the class mass of distribution contribute to the posterior $P(c_i|x)$ more than categories representing class outliers.

When imposing $P_{\min} = 1$ in (12), $\hat{P}(c_i|w_j)$ is forced to be either 1 or 0 and a decision about a class will be similar by both the GA and BA, i.e., the GA is a private case of the BA. $P_{\min} = 1$ also reduces the BA to the fast learning mode of the FA. To demonstrate this difference between the models, we illustrate in Fig. 1 an example two-class classification problem where the first class is associated with three categories and the second class with a single category. The figure shows that following the GA method of inference, patterns that should have been classified to class c_2 (e.g., the pattern indicated by an arrow) are wrongly classified to class c_1 only because the unweighted sum of joint probabilities is concerned. In contrast to the GA, the BA, weighting each joint probability for c_1 by the posterior probabilities $\hat{P}(c_1|w_j) < 1, j \in [1, 3]$ as computed on the training set, finds c_2 (having a single category weighted by $\hat{P}(c_2|w_4) = 1$) a more appropriate class for the pattern than c_1.

V. Experimentation

We compared the BA performance with respect to classification accuracy, learning curves, number of categories, sensitivity to class overlapping and risk with those of the FA and GA using synthetic and real-world databases. The FA was trained either for one epoch or until completion.

By using a diagonal (rather than full) covariance matrix for the cluster multidimensional Gaussian and making the mapping between a category and a class deterministic (rather than probabilistic), i.e., $P_{\min} = 1$ in (12) imposing $\hat{P}(c_i|w_j) = 0$ or 1, we degenerate the BA and render the model comparable to the GA. This allows us to evaluate the importance of the other differences between the models (Section IV and Table I). Also, there are two practical benefits in degenerating the BA. First is that

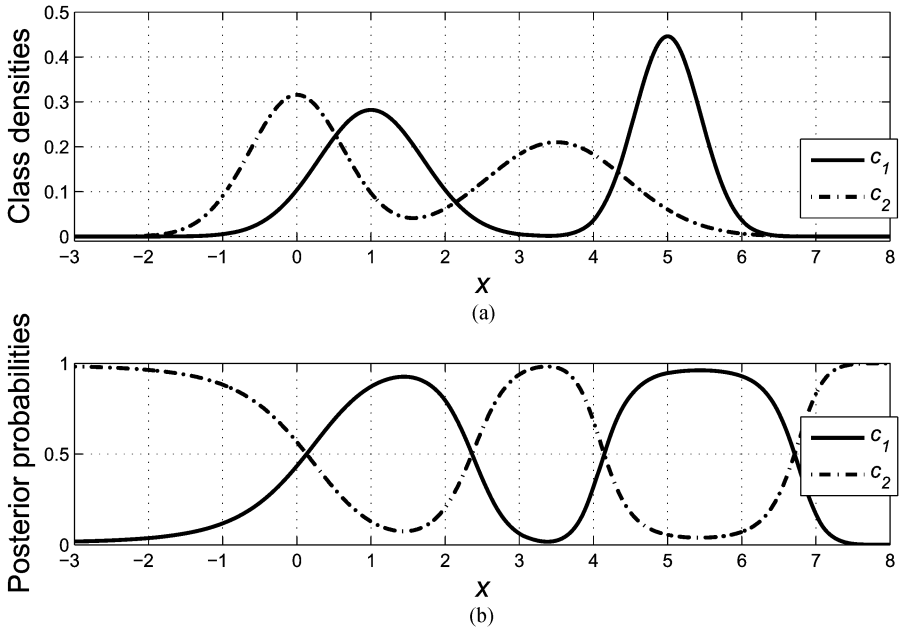

Fig. 2. (a) Generating class conditional probability densities. (b) Posterior probabilities for Experiments 1–3.

a diagonal matrix is a compromise between a spherical covariance, which can only model spherical densities, and a full covariance matrix, which can model any rotated scaled Gaussian density but requires too many parameters and has poor sequential estimate. Second, by setting $P_{\min} = 1$ there is only a single parameter (S_{MAX}) to optimize for the BA. Nevertheless, as has been outlined throughout the study the BA is general and not restricted to a diagonal matrix or deterministic class-category mappings.

A. Experimentation With Synthetic Data

We conducted five experiments with synthetic data formed using predefined generating models. The first four experiments were held employing Gaussian distributed data whereas the fifth experiment using non-Gaussian data. In the first experiment, we optimized the three classifiers to the data. In the second, we evaluated the classifiers learning curves. In the third experiment, we addressed the sensitivity of the classifiers to increasing degrees of statistical overlapping and, in the fourth, we introduced losses to different misclassifications. In all the experiments, the test accuracy was estimated using 2000 patterns and the training accuracy using different numbers of patterns depending on the experiment. Each experiment was performed using ten random replications of the data and the results were averaged.

The first three experiments were held using a simple 1-D classification task. Patterns were generated from two classes, each represented by a mixture of two Gaussian components

$$P(x|c_i) = \sum_{j=1}^{2} P_{ij} \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left\{-\frac{(x-\mu_{ij})^2}{2\sigma_{ij}}\right\}$$

$$\mu_{11} = 1 \quad \mu_{12} = 5 \quad \mu_{21} = 0 \quad \mu_{22} = 3.5$$

$$\sigma_{11} = 0.5 \quad \sigma_{12} = 0.2 \quad \sigma_{21} = 0.4 \quad \sigma_{22} = 0.9$$

$$P_{i1} = 0.5 \quad P_{i2} = 0.5 \quad i = 1, 2$$

$$P(c_1) = 0.7 \quad P(c_2) = 0.3.$$

The class conditional probability densities and class posterior probabilities for this classification task are shown in Fig. 2.

Experiment 1—Optimization: In the first experiment, the parameters of each classifier were optimized to the previously mentioned classification task. The FA vigilance and choice parameters were optimized by exhaustive search over a wide grid of parameters determined for $0 \leq \rho \leq 1$ and $10^{-12} \leq \alpha \leq 1$. The BA maximal category hypervolume parameter was optimized over the range $10^{-8} \leq S_{\mathrm{MAX}} \leq 10^{2}$. In the case of a diagonal covariance matrix, each variance in the initial covariance matrix should be very small, so that the category could grow and change its shape during training. We set the initial covariance matrix to be spherical $\Sigma_{\mathrm{init}} = \sigma_{\mathrm{init}}^2 I$, where I is the identity matrix, hence the parameter σ_{init}^2 is required to satisfy $\sigma_{\mathrm{init}}^2 \ll (S_{\mathrm{MAX}})^{1/M}$(8), (9) to assure that the initial category hypervolume is much smaller than the maximal hypervolume allowed. The γ parameter of the GA, i.e., the initial standard deviation of a category, was set to be the same as σ_{init} of the BA and the GA vigilance parameter ρ to 0.[10]

The optimal parameter values for each classifier, trained using 500 patterns, were determined based on the highest classification accuracy on a validation set of 2000 patterns independent of the training and test sets. Based on the accuracy averaged over the ten data replications, we selected $\rho = 0$ and $\alpha = 10^{-12}$ for an FA trained for one epoch (Fig. 3) and $\rho = 0.4$ and $\alpha = 10^{-9}$ for an FA trained until completion (Fig. 4).

Fig. 5 shows the BA training and validation accuracies, as well as the number of categories recorded for increasing (log) maximal category hypervolume values. The figure can be roughly divided into three different regions. The middle

[10]In the vigilance test of the GA (Table I), the log-likelihood (LL) is required to be larger than the vigilance parameter ρ in order to achieve category resonance [6]. Since the GA LL is nonpositive and usually $\rho \in [0, 1]$, we converted the LL to the likelihood by taking exp(LL) before conducting the test with $\rho = 0$ [20], providing minimal category size and thus maximal generalization ability to the model.

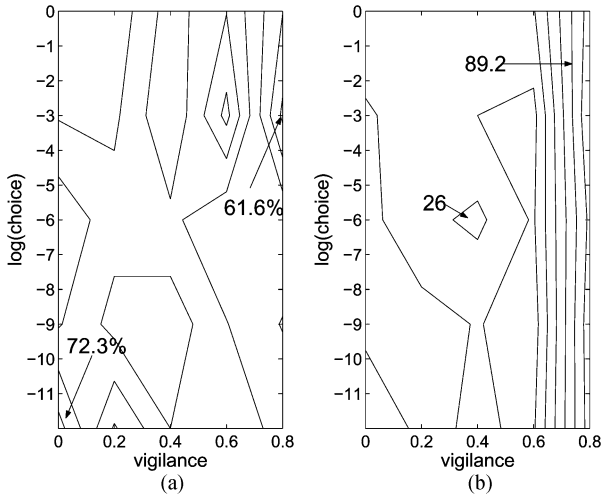

Fig. 3. (a) Test accuracy and (b) number of categories for various vigilance and choice parameter values of the FA trained for one epoch (Experiment 1).

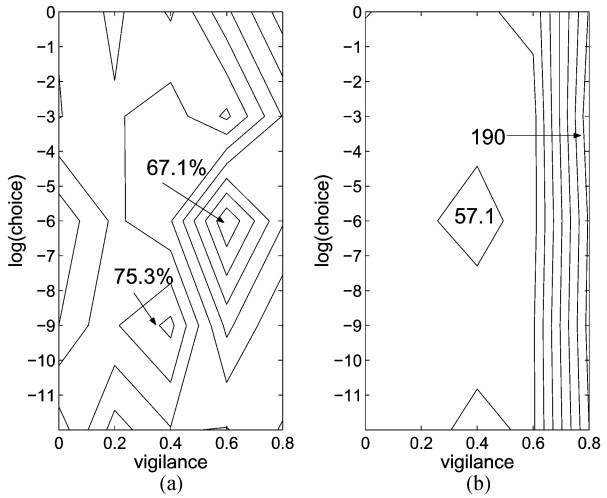

Fig. 4. (a) Test accuracy and (b) number of categories for various vigilance and choice parameter values of the FA trained until completion (Experiment 1).

region $(0.03 \leq S_{\mathrm{MAX}} \leq)1$, denoted as (1), yields the highest validation accuracy (maximum at 80.5%), almost fixed train accuracy (81.2%), and a relatively low number of categories (less than 39). The validation accuracy in this region is close to that of the Bayes' classifier (bound) [16] at 81.5%. The second region (2), defined for $S_{\mathrm{MAX}} > 1$, is characterized by a very low number of categories (lower than 11) and mostly constant train and validation accuracies (77.3% and 76.1%, respectively), which are lower than those of region 1. The third region (3) for $S_{\mathrm{MAX}} < 0.03$ yields decreasing validation accuracy (down to 70%), rising training accuracy (up to 100%) and increasing number of categories (up to 500) for decreasing maximal hypervolumes. That is, region 2 represents underfitting by the BA, region 3 represents overfitting, and region 1 is where the maximal category hypervolume parameter should be selected from. Based on the highest averaged validation accuracy achieved in region 1 of Fig. 5, we chose the value of the maximal category hypervolume parameter (S_{MAX}) to be 1.

Since in all the experiments with synthetic data the densities are 1-D (i.e., $M = 1$), we select the value of the initial parameter $(\sigma_{\mathrm{init}}^2 \ll 1^{1/1})$ to be two orders of magnitude lower than the maximal category hypervolume parameter, i.e., 0.01. This was also the value determined for the initial variance (γ) of the GA. When decreasing the maximal category variance to a low enough value, e.g., $S_{\mathrm{MAX}} = 10^{-8}$, the number of categories reaches the number of training patterns, i.e., the BA turns to be a Parzen window probability density estimator [kernel density estimator (KDE)] with a Gaussian kernel [17]. In addition, though the number of categories formed is random in nature, its mean forms a smooth function of the maximal variance that rises monotonically when lowering the maximal variance. Therefore and in order to lessen the computational load, we suggest to begin the optimization of this parameter with a high value of the maximal category hypervolume parameter corresponding to a small number of categories, lower this value, and use the classification accuracy on a validation set as a stopping criterion.

We note that the use of the full covariance matrix for each category of the BA has both advantages and risks. On one hand, using the full covariance might model each class more accurately than using the diagonal covariance, especially if the generating model has rotated clusters. On the other hand, the full covariance requires much more parameters than the diagonal one $(M(M+1)/2$ versus M for an M-dimensional feature space). Moreover, if a specific category clusters only a few training patterns, its full covariance matrix estimation will be poor which will undermine the classification accuracy. Thus, the full covariance matrix is useful for large databases or when using a relatively high maximal hypervolume parameter encouraging large categories (8) which can be well populated. However, as Fig. 5 reveals, high values of S_{MAX} do not contribute to high accuracy, so, eventually, we should restrict employing the full covariance matrix to large databases.

Finally for this experiment, Fig. 6 demonstrates the BA estimation for the posterior probability of class c_1 using the selected initial variance. For clarity, the estimation for c_2 (which is 1 minus that for c_1) is not shown in the figure. The estimation is good as long as the density for c_1 does not vanish, i.e., for $0 \leq x \leq\sim 6.5$ (see Fig. 2). Note that neither the FA nor the GA estimate posterior probabilities.

Experiment 2—Learning Curves: Using the same database, we investigated in this experiment the learning curves of the classifiers by measuring their test accuracy and number of categories as a function of the sample size. Using their optimal parameters (Experiment 1), the classifiers were trained on sets of increasing sizes (from 100 to 2000 patterns in increments of 100 patterns) and evaluated on the same test set. The advantage of the BA over all other classifiers with respect to both accuracy and number of categories and regardless of the sample size is demonstrated in Fig. 7. The BA accuracy is shown to be slightly inferior to the Bayes' bound estimated at 82.6% and superior to all other classifiers. The highest accuracies achieved by each of the classifiers are given in Table II, i.e., 82.3%, 81.5%, 73.5%, and 70.1% for the BA, GA, and FA trained until completion and FA trained for one epoch, respectively. The table also shows the category growth for the classifiers, quantifying

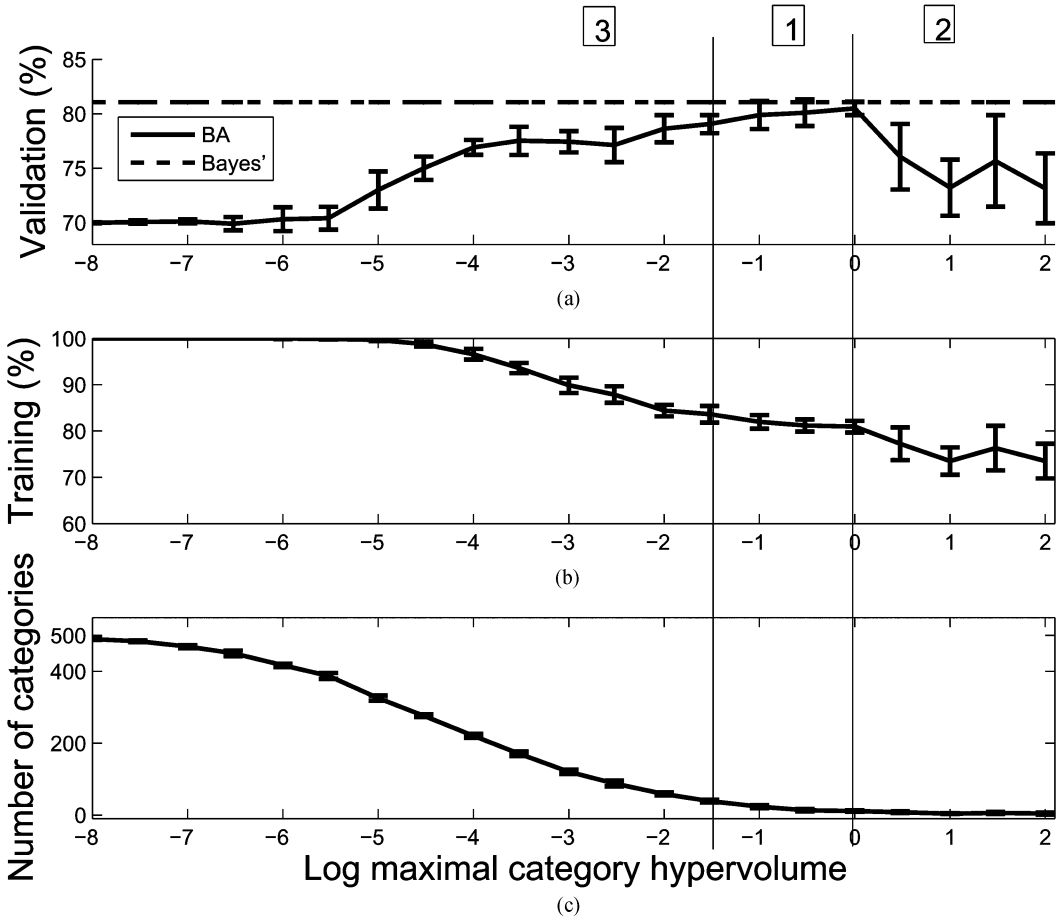

Fig. 5. (a) Validation accuracy, (b) training accuracy, and (c) number of categories of the BA for different values of the (log) maximal category hypervolume parameter (Experiment 1).

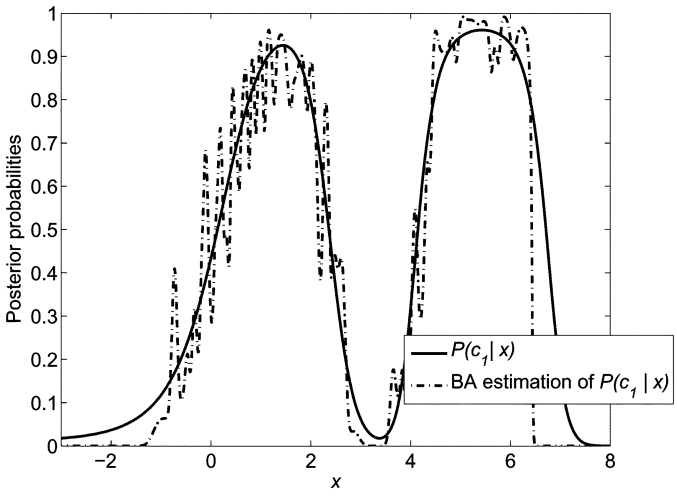

Fig. 6. Posterior class probability for c_1 and its estimation by the BA (Experiment 1).

the rate of increase of the number of categories with the increase of the sample size. Using linear regression [17], we computed values of category growth for the BA that are between an order and two orders of magnitude smaller than those of the GA and FA, respectively.

TABLE II
MAXIMAL TEST ACCURACY AND CATEGORY GROWTH RATE FOR THE FA TRAINED FOR ONE EPOCH OR UNTIL COMPLETION, GA, AND BA IN EXPERIMENT 2 (GAUSSIAN DATA)

Classifier	Test accuracy	Category growth
FA (one epoch)	70.1%	0.05
FA (until completion)	73.5%	0.16
GA	81.5%	0.05
BA	82.3%	0.007

Experiment 3—Statistical Overlapping: The purpose of this experiment was to investigate the sensitivity of the classifiers to statistical overlapping. We used the previous data-generating functions but changed the variance of each density component in order to control the statistical overlapping between the classes. The degree of overlapping was measured by the accuracy of the Bayes' classifier (upper bound on the classification accuracy). That is, as the variances of the densities increase and so does the degree of overlapping between classes, the accuracy of the Bayes' classifier reduces. We employed 500 and 2000 patterns for training and test, respectively. As can be seen in Fig. 8, all the classifiers produce good results (accuracy and number of categories) when the Bayes' bound is very high (95%-100%), i.e., almost no statistical overlapping between the classes. When

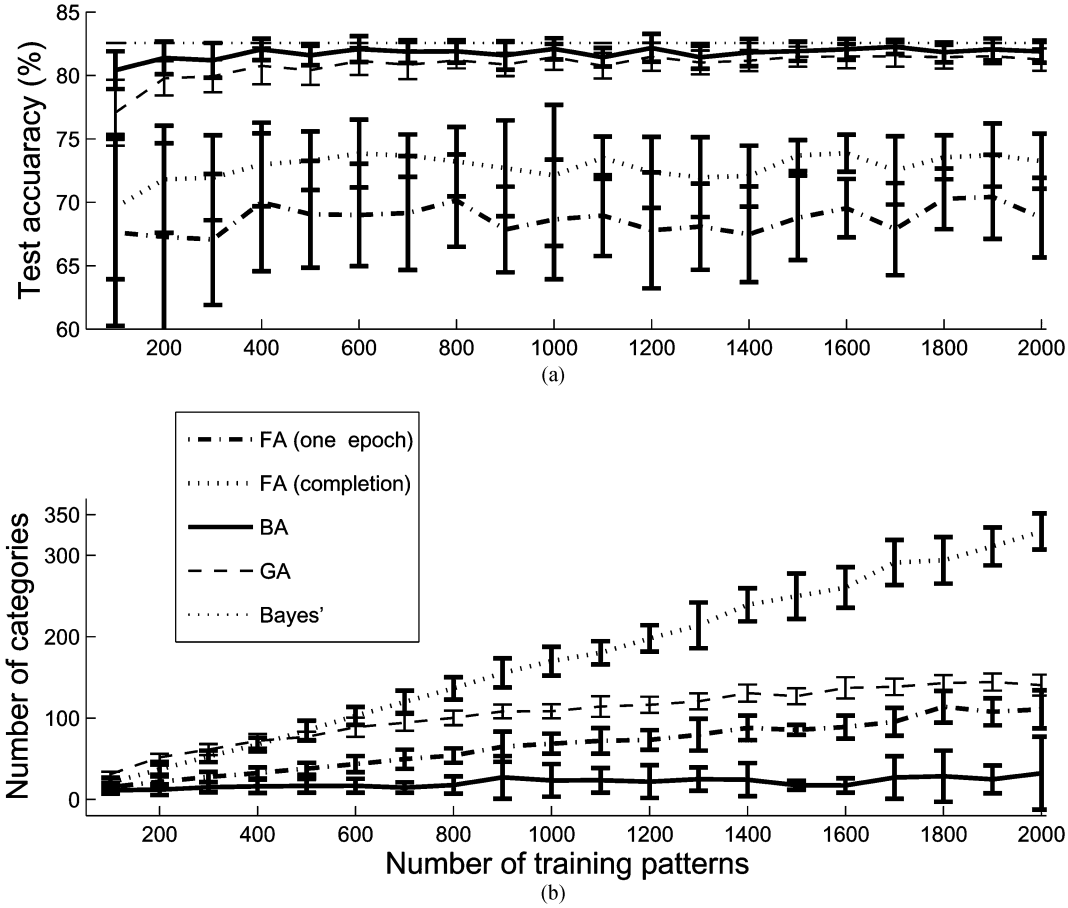

Fig. 7. (a) Test classification accuracy and (b) number of categories for the FA trained for one epoch or until completion, GA, and BA for increasing sample sizes in Experiment 2 (Gaussian data). The accuracies are also compared to that of the Bayes' classifier (bound).

lowering the Bayes' bound (i.e., increasing the degree of overlapping), the BA and GA average accuracies remain close to the Bayes' bound that is represented by the line $y = x$. The accuracies of the BA (71.2%) and GA (69.6%) for the lowest Bayes' bound that is measured are close to that of the Bayes' bound (74.3%) where those of the FA trained for either one epoch or until completion are 58.6% or 62.7%, respectively. The advantage of the BA over the other classifiers with respect to the number of categories is even clearer. The FA trained until completion, FA trained for one epoch, and GA require 363.4, 185, and 108.6 categories, respectively, on average for the lowest Bayes' bound measured compared to only 41.7 categories on average for the BA.

Experiment 4—Risk: We evaluated the classifiers for the case where misclassifications have different costs. We adopt the known problem of an animal that has to distinguish between a predator and a harmless animal in order to decide whether to run away or stay. A simplified version of this problem can be formalized easily using the statistical framework when defining two classes. The first class—harmless—has a high *a priori* probability (say 0.9, as it is typical to encounter a harmless animal) and the second class—predator—has a low *a priori* probability (0.1). Let us assume that the predator and harmless classes have a similar feature leading to statistical overlapping between the classes. The cost of misclassifying the predator is established as 50, as it can cause the death of the animal, where the cost of misclassifying a harmless animal (false alarm) is low (say 1), as it only forces the animal to run and hide for a short period. This defines the loss function $\lambda(c_i|c_j)$, i.e., the loss in deciding on c_i where the true class is c_j. Thus, we cast our example problem as follows:

$$P(x|\text{Harmless}) = \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma_{11}^2}}\exp\left\{-\frac{(x-\mu_{11})^2}{2\sigma_{11}^2}\right\} + \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma_{12}^2}}\exp\left\{-\frac{(x-\mu_{12})^2}{2\sigma_{12}^2}\right\}$$

$$P(x|\text{Predator}) = \frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}$$

$$\mu_{11} = 1 \quad \mu_{12} = 3 \quad \mu_2 = 2$$

$$\sigma_{11} = 0.4 \quad \sigma_{12} = 0.4 \quad \sigma_2 = 0.15$$

$$P_{\text{harmless}} = 0.9 \quad P_{\text{predator}} = 0.1$$

$$\lambda(c_i|c_j) = \begin{bmatrix} 0 & 50 \\ 1 & 0 \end{bmatrix}.$$

The class-conditional probability density and conditional loss (risk) [16]

$$R(c_i|x) = \sum_{j=1}^{2} \lambda(c_i|c_j)\hat{P}(c_j|x) \tag{21}$$

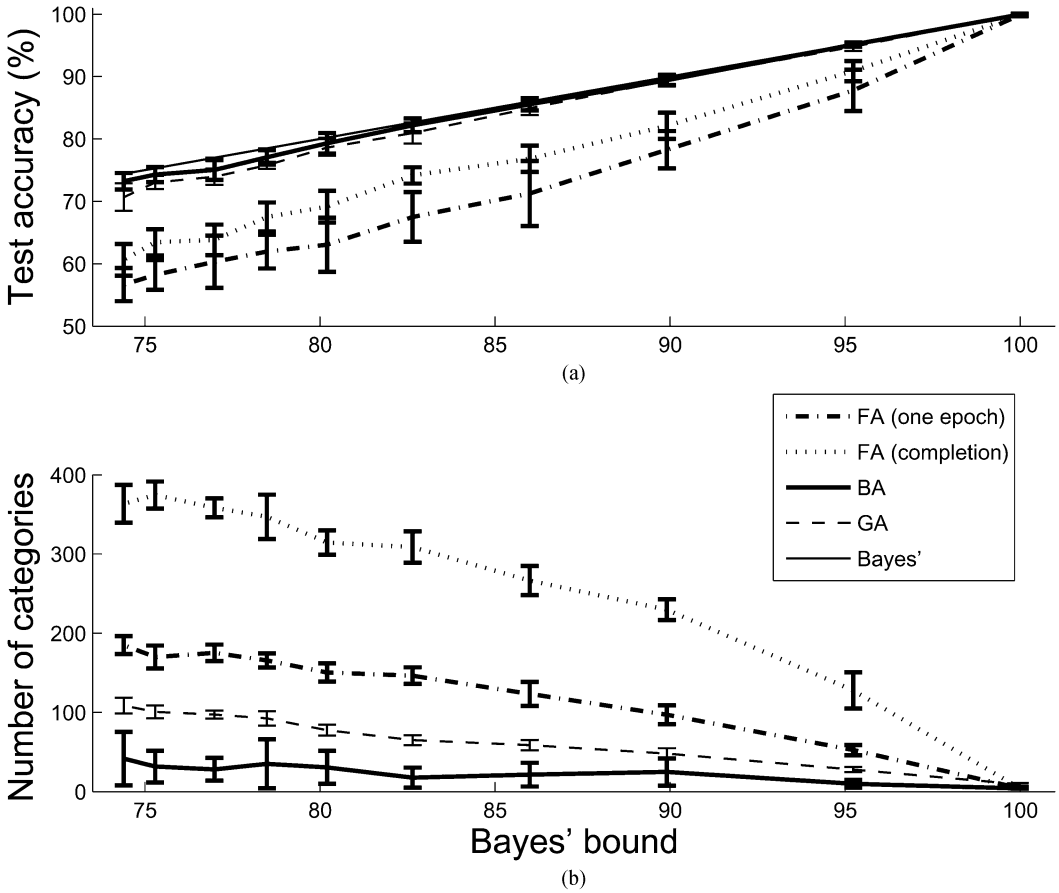

Fig. 8. (a) Test classification accuracy and (b) number of categories for the FA trained for one epoch or until completion, GA, and BA for increasing Bayes' bound values (decreasing statistical overlapping) (Experiment 3). Accuracies are also compared to that of the Bayes' classifier (bound).

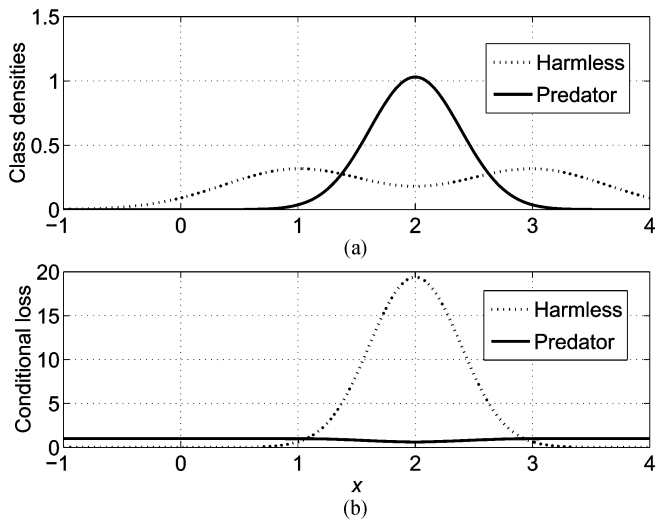

Fig. 9. (a) Class-conditional probability densities and (b) conditional losses for the two classes of Experiment 4.

for each class are shown in Fig. 9. Both classifiers were optimized (similarly to the optimization in Experiment 1), trained, and tested using the previously described densities. Having a set of N pairs $(x_1, y_1), \ldots (x_N, y_N)$, where x_n is a pattern and y_n is its label, as well as a decision rule $\alpha(x_n)$, the overall risk, which is the expected loss associated with this decision rule, is [16]

$$R(\alpha) = \frac{1}{N} \sum_{n=1}^{N} \lambda(\alpha(x_n)|y_n). \tag{22}$$

The expected loss and number of categories for sets of increasing sizes (from 100 to 2000 patterns in increments of 100 patterns) for both classifiers are presented in Fig. 10. In order to minimize the expected loss (22), we have to minimize the conditional loss (21) for each pattern, i.e., select class $i^* = \arg\min_i R(c_i|x),\ \forall\, x$ [16]. The BA estimating the posterior probabilities (19) of the two classes can easily implement this Bayes' decision rule, however not the GA utilizing likelihoods rather than posteriors or the FA obtaining only a Boolean output in fast learning. In order to provide the GA the ability to use the conditional loss, and thereby, reduce its expected loss, we estimated for the GA a posterior probability for each class using the ratio of the sum of likelihoods of all categories associated with the class and the sum of likelihoods of all categories. Indeed, this suggestion enabled the GA to obtain an expected loss similar to that of the BA and the Bayes' expected loss outperforming the large expected losses of the two FA variants (see Fig. 10). The minimal expected loss of each classifier is shown in Table III in comparison to the Bayes' bound of 0.47. The differences between the models estimating

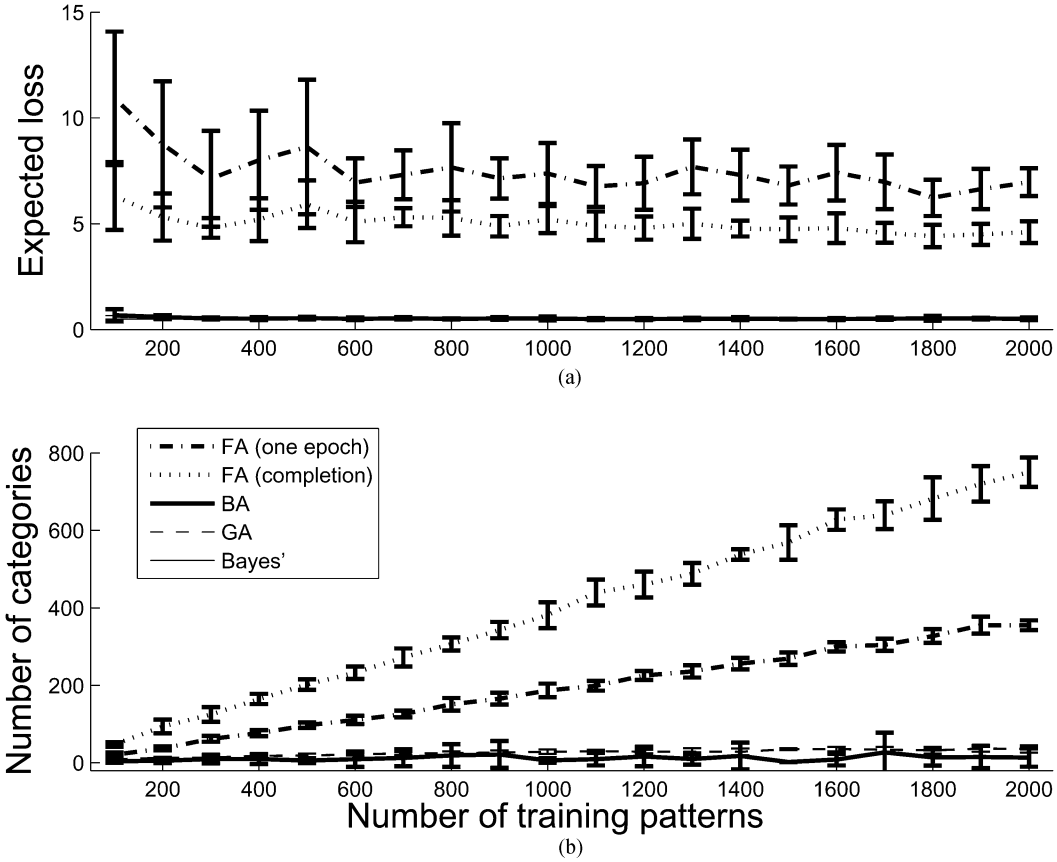

Fig. 10. (a) Expected loss and (b) number of categories of the FA trained for one epoch or until completion, GA, and BA for increasing sample sizes (Experiment 4). Also shown is the expected loss of the Bayes' classifier.

TABLE III
MINIMAL EXPECTED LOSS AND CATEGORY GROWTH RATE FOR THE FA TRAINED FOR ONE EPOCH OR UNTIL COMPLETION, GA, AND BA IN EXPERIMENT 4

Classifier	Minimal expected loss	Category growth
FA (one epoch)	6.22	0.18
FA (until completion)	4.41	0.37
GA	0.5	0.013
BA	0.49	0.004

the posterior probability (i.e., the BA and, employing the previous suggestion, also the GA) and those which are not (the FAs) are evident. Also evident (Table III) is the superiority of the BA to all other classifiers with respect to category growth.

Experiment 5—Non-Gaussian Densities: This experiment is very similar to Experiment 2 (learning curves); however, the classes are composed of non-Gaussian densities. The first class is composed of a mixture of a uniform and Rayleigh densities and the second class is a mixture of two uniform densities. The class-conditional densities and posterior probabilities for the two classes are shown in Fig. 11.

After all classifiers have been optimized to the problem (as in Experiment 1), the BA produced test accuracy which was only slightly lower than the Bayes' bound (85.7%) when measured for increasing sample sizes (Fig. 12 and Table IV). This accuracy is almost insensitive to the sample size. The GA accuracy,

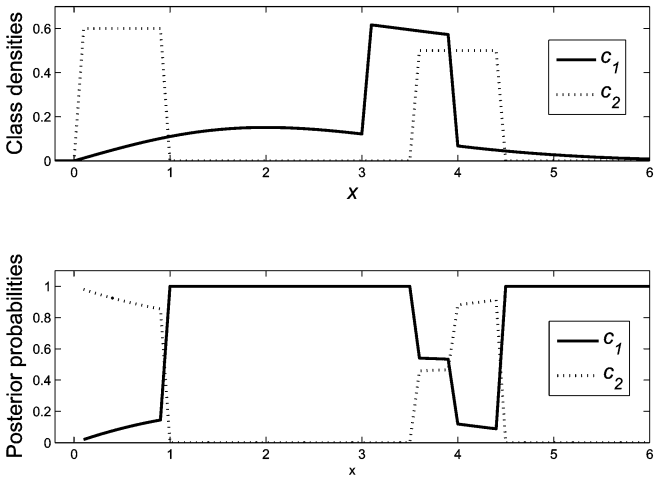

Fig. 11. (a) Generating (non-Gaussian) class-conditional densities and (b) posterior probabilities for the two classes of Experiment 5.

lower than that of the BA in $\sim 2-7\%$, was more sensitive to the sample size. The FA accuracy was lower than that of the BA in about 10% for both training modes. Thus, the high test accuracy attained by the BA demonstrates its superior ability to approximate nondifferentiable and non-Gaussian densities. This superiority is also shown in the category growth of the BA which is lower in at least an order of magnitude than those of the other

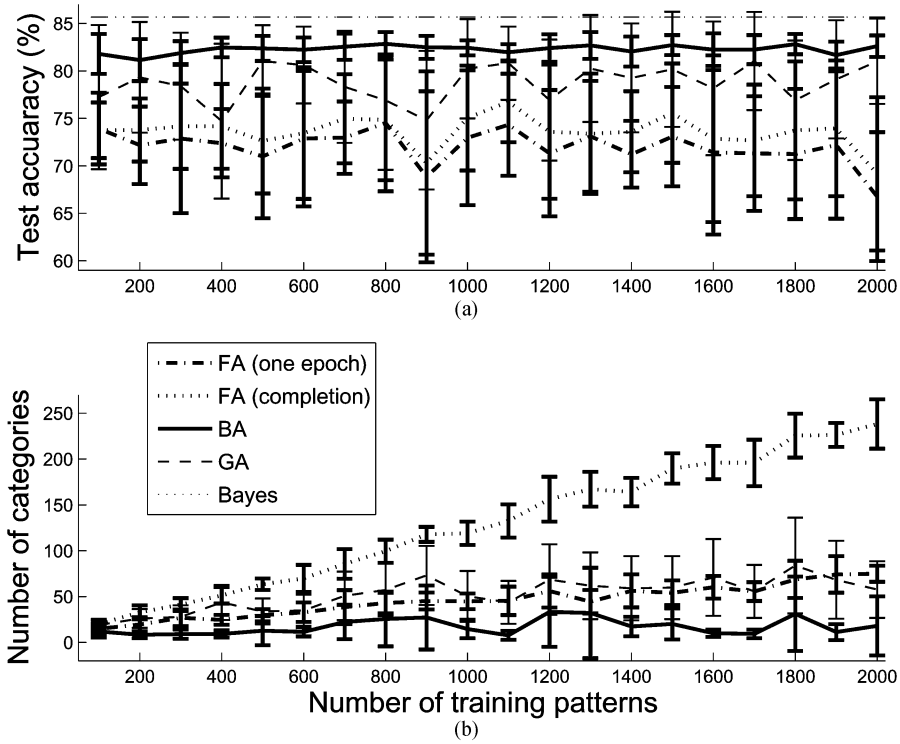

Fig. 12. (a) Test classification accuracy and (b) number of categories for the FA trained for one epoch or until completion, GA, and BA for increasing sample sizes in Experiment 5 (non-Gaussian data). Accuracies are also compared to that of the Bayes' classifier (bound).

TABLE IV
MAXIMAL TEST ACCURACY AND CATEGORY GROWTH RATE FOR THE FA TRAINED FOR ONE EPOCH OR UNTIL COMPLETION, GA, AND BA IN EXPERIMENT 5 (NON-GAUSSIAN DATA)

Classifier	Test accuracy	Category growth
FA (one epoch)	74.5%	0.03
FA (until completion)	76.7%	0.12
GA	81.0%	0.02
BA	82.9%	0.004

classifiers, similarly to all previous experiments. In addition, examining the standard deviations with respect to both accuracy and number of categories, the BA is more stable and reliable than all other classifiers.

B. Experimentation With Real-World Databases

The BA was investigated using real-world classification problems from the University of California at Irvine (UCI) repository [18], United-States Postal Service (USPS) database [19] containing segmented handwritten digits from U.S. zip codes, and a cytogenetic database [21] in which the detection of signals representing Down syndrome, Patau syndrome, or both syndromes[11] enable genetic abnormality diagnosis [11]. The BA was compared to the FA trained for one epoch or until completion, a single Gaussian estimation (SGE) classifier, and the kernel density estimation (KDE) classifier [17]. The KDE classifier is an extreme case of the BA for a very small maximal hypervolume S_{MAX} and the SGE classifier that uses a diagonal covariance matrix is a computationally efficient algorithm.

[11]For each syndrome, we distinguish between real signals and artifacts.

The GA was not included in the comparison on the real-world databases for two reasons. First is as demonstrated in Section IV—the GA is a private case of the BA so there is no point in experimenting with both models. Indeed, in the experiments with the synthetic data (Section V-A), the performances (accuracy or risk) of the GA were always inferior to those of the BA and the number of categories was always higher than that of the BA. Second, no method is suggested in [6] or elsewhere for optimizing the GA parameters—ρ and γ. For the synthetic databases, we adopted the BA σ^2_{init} for the GA γ parameter (as they are used for the same purpose) and the recommendation in [6], [20] to set $\rho = 0$. However, choosing these values may not be justified to either of the databases tested in this section and the GA may therefore operate suboptimally.

The experimentation method used was cross validation [17] using CV10. The optimization of the FA and BA for each of the 20 databases tested was similar to the optimization performed for the synthetic databases. All experiments were performed using the averaging strategy [1] with five different data presentation orders. The same ordered training sets and CV folds were used for the FA and BA to ensure proper comparison. The test average classification accuracy and number of categories for the classifiers are presented in Tables V and VI, respectively. Bold font emphasizes the highest accuracy (Table V) or minimal number of categories (Table VI) achieved for a database. Missing entries for the FA trained until completion and two large databases (Pen and USPS) (NA in Tables V and VI) are due to exceeding memory and computation resources resulting in simulation collapse of the FA, as hundreds and thousands of categories were formed.

When comparing the BA and FA trained for one epoch (Table V), the former is more accurate on 17 databases and the

TABLE V
TEST ACCURACIES (%) OF THE FA TRAINED FOR ONE EPOCH OR UNTIL COMPLETION, BA, SGE, AND KDE CLASSIFIERS ON 20 REAL-WORLD DATABASES. THE "*" INDICATES DATABASES WHERE THE ADVANTAGE OF THE BA OR FA TRAINED UNTIL COMPLETION OVER THE OTHER CLASSIFIER IS STATISTICALLY SIGNIFICANT WITH SIGNIFICANCE LEVEL OF 0.05

Database	BA	FA (one epoch)	FA (until completion)	SGE	KDE
Ionosphere	87.4 (5.7)	84.7 (7.1)	**89.2 (5.3)**	82.0 (5.7)	79.8 (5.9)
Wine	93.6 (5.6)	92.3 (6.8)	92.6 (5.7)	**95.3 (4.9)**	87.2 (9.1)
Iris	**95.7 (4.4)** *	91.7 (8.9)	90.3 (10.2)	95.3 (5.7)	93.2 (6.8)
Glass	**64.7 (9.6)** *	45.6 (15.2)	45.6 (15.2)	46.9 (10.6)	32.8 (11.2)
Image	**84.6 (8.0)** *	78.0 (14.2)	77.9 (14.1)	78.3 (9.5)	19.6 (8.0)
Liver	**61.2 (8.5)**	59.6 (7.5)	58.6 (9.1)	55.7 (8.0)	61.1 (6.3)
Pima	72.8 (6.5) *	64.4 (6.7)	66.9 (6.0)	**75.6 (4.5)**	65.1 (5.5)
Breast	95.8 (2.1)	94.8 (2.7)	95.2 (2.7)	**96.1 (2.3)**	95.6 (2.1)
Down syndrome	**87.1 (3.2)**	83.9 (3.3)	85.8 (2.8)	80.0 (3.7)	69.2 (3.2)
Patau syndrome	84.6 (3.6)	79.2 (3.8)	84.4 (2.8)	**86.6 (2.5)**	57.9 (4.1)
Down & Patau (4-class)	80.1 (2.7)	79.0 (2.2)	**84.5 (1.9)** *	74.5 (2.2)	68.0 (2.9)
Balance	81.7 (6.3) *	75.8 (5.5)	76.4 (5.1)	**90.3 (3.9)**	89.9 (3.6)
Car	89.6 (2.9)	93.3 (2.1)	**95.7 (1.5)** *	80.3 (3.1)	77.5 (2.9)
Mushroom	94.3 (3.5)	99.4 (1.5)	99.4 (1.5) *	87.4 (1.0)	**100.0 (0)**
Krkp	86.7 (2.9)	82.9 (4.0)	**89.4 (3.4)** *	52.2 (3.0)	77.2 (3.2)
Zoo	**94.9 (9.4)** *	47.9 (28.9)	47.9 (28.9)	40.6 (14.1)	83.7 (10.6)
Lymphography	**79.5 (9.5)**	76.9 (15.7)	76.9 (15.6)	78.2 (9.5)	78.7 (9.9)
Hayes	**71.8 (13.2)** *	65.5 (13.8)	65.5 (17.3)	68.9 (12.0)	63.0 (11.1)
Pen	95.7 (0.5)	96.7 (0.6)	NA	76.0 (1.4)	**99.3 (0.2)**
USPS	94.0 (0.4)	87.1 (1.4)	NA	90.8 (0.9)	**96.5 (0.7)**

latter on three databases. The BA requires a smaller number of categories on ten databases and the FA trained for one epoch on ten other databases (Table VI). Compared to the FA trained until completion, the BA accuracy is superior on 13 databases, inferior on five databases, and on another two databases the FA did not complete the task. Also, Table VI shows that the BA requires a smaller number of categories on 14 databases and the FA on four databases. In addition, we indicate in Table V if there is statistical significance to the superiority of the BA (FA trained until completion) to the FA trained until completion (BA). "*" denotes that the advantage of the classifier having a larger accuracy is statistically significant with significance level of 0.05. From the 13 (5) databases on which the BA (FA) has an advantage over the FA (BA), on 7 (4) databases it is statistically significant. Note that we exclude from consideration the two databases (Pen and USPS) on which the FA collapsed.

On average, the BA has superior test accuracy to the KDE and SGE classifiers. However, on several databases (Wine, Pima, Breast, Balance, and Patau syndrome), the SGE classifier yields higher test accuracy due to either a small sample size, almost normally distributed data, or both, which are factors in favor of the (parametric) SGE. The KDE classifier yields test accuracy higher than the BA on the Balance, Mushroom, Pen, and USPS databases. As the KDE classifier is an extreme case of the BA, the results show that on these four databases the optimization procedure suggested for the maximal hypervolume parameter (Experiment 1 in Section V-A) stops too early at local maxima (81.7%, 94.3%, 95.7%, and 94.0% for these databases, respectively) and misses higher maxima reached by the KDE classifier (at 89.9%, 100%, 99.3%, and 96.5%, respectively).

Averaging the results over all databases shows (Table VII) that the BA outperforms all other classifiers with respect to accuracy. The second best classifier, the FA trained until completion, has accuracy which is lower by 4.9% than that of the BA. In addition, the BA is more robust than the FA (using either of the training modes), as its accuracy variance is smaller (6% compared to 8.3%). Moreover, when classifying only noisy databases (i.e., Glass, Image, Lymphography, Zoo, and Hayes), the BA produces relatively stable classification results, demonstrated in an average variance over these databases of 9.9%, where the FA trained for one epoch or until completion have variances of 17.6% and 18.2%, respectively. Finally, the number of BA categories averaged over all 20 databases is much closer to that of the FA trained for one epoch than to that of the FA trained until completion.

VI. DISCUSSION

We modified the FA using the Bayesian framework in order to enhance the model classification accuracy while simultaneously reduce its category proliferation. The proposed BA preserves the FA advantages and also improves the latter performance

TABLE VI
NUMBER OF CATEGORIES FORMED BY THE FA TRAINED FOR ONE EPOCH OR UNTIL COMPLETION AND THE BA FOR 20 REAL-WORLD DATABASES

Database	BA	FA (one epoch)	FA (until completion)
Ionosphere	19.3 (1.9)	**16.8 (2.2)**	25.0 (2.1)
Wine	8.7 (1.3)	**7.8 (1.1)**	9.8 (2.7)
Iris	**7.3 (0.7)**	8.6 (1.5)	19.2 (4.8)
Glass	55.6 (2.8)	**30.0 (3.7)**	31.7 (3.6)
Image	31.7 (1.6)	**20.3 (2.7)**	20.8 (2.7)
Liver	40.8 (1.7)	**15.8 (3.2)**	94.1 (6.3)
Pima	**10.7 (1.9)**	57.9 (5.9)	177.2 (12.8)
Breast	**5.9 (1.5)**	8.8 (1.5)	13.4 (2.5)
Down syndrome	58.7 (3.8)	**45.5 (7.0)**	103.7 (7.2)
Patau syndrome	178 (18.4)	**98.1 (5.6)**	122.5 (4.9)
Down & Patau (4-class)	311 (17.1)	**205.5 (6.4)**	236.3 (9.1)
Balance	**16.9 (6.1)**	85.0 (10.8)	111.0 (6.9)
Car	**59.4 (20.9)**	60.5 (7.22)	78.8 (6.4)
Mushroom	**8.6 (1.6)**	12.9 (1.5)	12.9 (1.5)
Krkp	127.5 (32.8)	**65.8 (4.6)**	182.6 (18.7)
Zoo	**9.4 (0.9)**	20.8 (1.4)	20.8 (1.4)
Lymphgraphy	**11.7 (2.7)**	19.5 (4.5)	19.6 (4.3)
Hayes	**21.9 (5.3)**	43.6 (5.4)	59.8 (5.2)
Pen	**66.8 (7.5)**	128.6 (9.3)	NA
USPS	152.3 (6.8)	**143 (7.6)**	NA

TABLE VII
TEST CLASSIFICATION ACCURACY AND ACCURACY VARIANCE, AS WELL AS NUMBER OF CATEGORIES FOR THE BA, FA TRAINED FOR ONE EPOCH OR UNTIL COMPLETION, SGE, AND KDE CLASSIFIERS AVERAGED OVER ALL THE 20 REAL-WORLD DATABASES STUDIED

Classifier	Test accuracy (%)	Test variance (%)	Number of categories
BA	84.8	6.0	54.6
FA (until completion)	79.9	8.3	74.4
FA (one epoch)	78.5	8.3	45.7
Single Gaussian estimation classifier	76.5	5.9	NA
Kernel density estimation classifier	74.8	5.9	NA

by the following: 1) representing categories using multidimensional Gaussian distributions, 2) allowing categories to grow or shrink, 3) limiting the volume of categories, 4) employing the Bayes' decision theory to probabilistically associate patterns to categories and categories to classes, thereby augmenting both learning and inference, and 5) predicting a class using all the categories associated with this class.

Indeed, the BA outperformed the FA, either trained for one epoch or until completion, for almost every aspect of performance—classification accuracy, learning curves, sensitivity to class overlapping, expected loss, and number of categories, on both synthetic and real-world data. On the synthetic databases, the dominance of the BA was absolute. This dominance was also evident to the GA. In all the experiments with the synthetic data, the accuracy of the BA was very close, and sometimes even identical, to the Bayes' bound of accuracy. In addition, averaged over all the real-world databases, the BA accuracy outperformed those of the FA classifiers. On all the real-world databases for which the FA trained for one epoch needed a smaller number of categories than the BA, it was also less accurate than the BA. On three out of the four databases on which the FA trained until completion was more accurate than the BA, it also needed at least 30% more categories than the BA. Also, the FA produced less stable and reliable classification accuracy than the BA.

The results on the synthetic databases show that the FA trained until completion suffers from category proliferation the most, both the FA trained for one epoch and GA suffer from moderate category proliferation, and the BA is the least sensitive classifier to the increase in data complexity. Note that by setting the GA vigilance parameter $\rho \equiv 0$ following [6], we form the minimal number of categories for this model, and thereby, provide the model of the maximal generalization capability [13]. That is, any optimization made to this parameter in order to improve the GA accuracy will also cause the increase in the number of categories formed which will intensify the GA inferiority to the BA with respect to category proliferation.

On the real-world databases, the BA number of categories was roughly comparable to that of the FA trained for one epoch, as each classifier needed to the smallest number of categories on ten other databases. However, averaged on all databases, the FA trained for one epoch required less categories. These two classifiers utilized fewer categories than the FA trained until completion. Although using the Bayesian framework, the BA number of categories remained constant only in special, not too complex, cases. This is since all the FA-based methods create new categories, but do not remove or join categories. Therefore, the number of categories in these models can only grow. The task of removing categories efficiently is not simple in the FA but can be performed elegantly by the BA using the category prior estimates $\hat{P}(w_j)$. Clearly, categories with low prior probabilities have little influence in the inference stage and they can safely be removed.

In addition to its improved performance, the BA estimates the class and category posterior probabilities in the ARTMAP and ART stages, respectively, while the FA and GA just propose a class (ARTMAP) or category (ART). By using class posterior probabilities (in a generative model such as the BA) in comparison to just making decisions (in a discriminative model such as the GA or FA), we can accommodate different priors between the training and test sets, e.g., in medical diagnosis of a rare disease [17]. In addition, we can address loss and classify according to the minimum expected loss, as was indeed performed in this study (Experiment 4 in Section V-A). Also, we may decide to reject a pattern if the maximal conditional loss (21) (or the posterior itself) is greater (smaller) than a threshold. These benefits are only possible when the classifier computes posterior probabilities. Usually, however, the accuracy of a generative model is more sensitive to the sample size than that of a discriminative model, since the former model requires estimating densities rather than just making decisions. However, this is not the case with the BA that manifests high accuracy, stable learning curves, and a small, constant number of categories even with a small sample size.

Finally, a way to overcome over the FA sensitivity to the order of presentation of training patterns and improve generalization performance is the voting strategy [1], [6], [9], [11]. It will be interesting to see whether the BA also improves its generalization performance using the voting strategy, although as stated before the BA is less sensitive to the order of data presentation than the FA. Another direction of future research is studying the contribution of the probabilistic association of categories to classes as revealed in $\hat{P}(c_i|w_j)$. Rather than degenerating the BA to be comparable to the GA by taking $P_{\min} = 1$, we plan to investigate the contribution of the probabilistic association between categories and classes to inference.

REFERENCES

[1] G. A. Carpenter, S. Grossberg, N. Markuzon, J. Reynolds, and D. Rosen, "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps," *IEEE Trans. Neural Netw.*, vol. 3, no. 5, pp. 698–713, Sep. 1992.

[2] S. Grossberg, "Adaptive pattern recognition and universal encoding II: Feedback, expectation, olfaction, and illusions," *Biol. Cybern.*, vol. 23, pp. 187–202, 1976.

[3] G. A. Carpenter, S. Grossberg, and J. H. Reynolds, "ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network," *Neural Netw.*, vol. 4, pp. 565–588, 1991.

[4] S. Marriott and R. F. Harrison, "A modified fuzzy ARTMAP architecture for the approximation of noisy mappings," *Neural Netw.*, vol. 8, pp. 619–641, 1995.

[5] M. A. Rubin, "Application of fuzzy ARTMAP and ART-EMAP to automatic target recognition using radar range profiles," *Neural Netw.*, vol. 8, pp. 1109–1116, 1995.

[6] J. R. Williamson, "Gaussian ARTMAP: A neural network for fast incremental learning of noisy multidimensional maps," *Neural Netw.*, vol. 9, pp. 881–897, 1996.

[7] E. G. Sánchez, Y. A. Dimitriadis, J. M. Cano-Izquierdo, and J. López-Coronado, "μ -ARTMAP: Use of mutual information for category reduction in fuzzy ARTMAP," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 58–69, Jan. 2002.

[8] Y. Suzuki, "Self-organizing QRS-wave recognition in ECG using neural networks," *IEEE Trans. Neural Netw.*, vol. 6, no. 6, pp. 1469–1477, Nov. 1995.

[9] G. A. Carpenter and N. Markuzon, "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases," *Neural Netw.*, vol. 11, pp. 323–336, 1998.

[10] G. A. Carpenter and W. D. Ross, "ART-EMAP: A neural network architecture for object recognition by evidence accumulation," *IEEE Trans. Neural Netw.*, vol. 6, no. 4, pp. 805–818, Jul. 1995.

[11] B. Vigdor and B. Lerner, "Accurate and fast off and online fuzzy ARTMAP-based image classification with application to genetic abnormality diagnosis," *IEEE Trans. Neural Netw.*, vol. 17, no. 5, pp. 1288–1300, Sep. 2006.

[12] A. Koufakou, M. Georgiopoulos, G. Anagnostopoulos, and T. Kasparis, "Cross-validation in fuzzy ARTMAP for large databases," *Neural Netw.*, vol. 14, pp. 1279–1291, 2001.

[13] I. Dagher, M. Georgiopoulos, G. Heilman, and G. Bebis, "An ordering algorithm for pattern presentation in fuzzy ARTMAP that tends to improve generalization performance," *IEEE Trans. Neural Netw.*, vol. 10, no. 4, pp. 768–778, Jul. 1999.

[14] G. A. Carpenter, S. Grossberg, and D. Rosen, "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Netw.*, vol. 4, pp. 759–771, 1991.

[15] G. Heilman and G. B. Georgiopoulos, "Order of search in fuzzy ART and fuzzy ARTMAP: Effect of the choice parameter," *Neural Netw.*, vol. 9, pp. 1541–1559, 1996.

[16] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[17] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[18] C. B. S. Hettich and C. Merz, "UCI Repository of Machine Learning Databases," 1998 [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[19] US Postal Service Handwritten Digits Recognition Corpus [Online]. Available: http://www.cedar.buffalo.edu/Databases/CDROM1

[20] J. R. Williamson, *private communication*.

[21] B. Lerner, W. F. Clocksin, S. Dhanjal, M. A. Hultén, and C. M. Bishop, "Feature representation and signal classification in fluorescence in-situ hybridization image analysis," *IEEE Trans. Syst., Man, Cybern. A, Syst, Humans*, vol. 31, no. 6, pp. 655–665, Nov. 2001.

Boaz Vigdor received the B.Sc. and M.Sc. degrees from Ben-Gurion University, Beer-Sheva, Israel, in 2002 and 2005, respectively, where currently, he is working toward the Ph.D. degree.

Boaz Lerner (M'07) received the B.A. degree in physics and mathematics from the Hebrew University, Israel, in 1982 and the Ph.D. degree in computer engineering from Ben-Gurion University, Beer-Sheva, Israel, in 1996.

He performed research at the Neural Computing Research Group, Aston University, Birmingham, U.K., and the Computer Laboratory of the University of Cambridge, Cambridge, U.K. In 2000, he joined the Department of Electrical and Computer Engineering at Ben-Gurion University, where currently, he is 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.