Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

# Bayesian Polytope ARTMAP: An ART-based network with two kinds of inner geometry categories

Leonardo Liao, Yongqiang Wu
Southwest Research Institute of Electronics and Telecommunication Technology
Chengdu 610041, P.R China
E-mail: liaoxp@mail.ustc.edu.cn

***Abstract*—The ART-based neural networks summarize data into groups via the use of inner categories. A category's template elements are updated incrementally in the light of new evidence provided by the presentation of input patterns. In order to reduce approximation error, this paper proposes Bayesian Polytope ARTMAP (BPTAM) which incorporates both simplex categories and Gaussian categories. During training, the simplex categories expand only towards the input pattern without category overlap, while the Gaussian categories grow or shrink by limiting their hypervolumes. In addition, BPTAM uses Bayes' decision theory for learning and inference, which makes BPTAM robust to noise and category overlap. Based on some preliminary but illustrative experimental results, BPTAM shows better applicability to data sets with noise, statistical overlapping and irregular geometry.**

***Keywords*—inner geometry category, Bayes' decision theory, generalization capability, classification, ART-based network.**

## I. INTRODUCTION

Adaptive resonance theory (ART) and ARTMAP neural networks, originally proposed by Grossberg, Carpenter and other researchers at Boston University [1]-[4], have proven to be very effective in many areas of pattern recognition. The fuzzy ARTMAP (FA) appears as one of the leading neural network (NN) algorithms for classification tasks. There are two aspects of FA should be noted. (1) The major drawback of the FA is its sensitivity to noise and statistical overlapping between the classes, which can give rise to category proliferation. In order to tackle this drawback which is an issue of approximation error [5], researchers have proposed several modifications to the FA such as Boosted ARTMAP (BARTMAP) [6], Micro-ARTMAP (μARTMAP) [7], Hierarchical ARTMAP (HARTMAP) [8], ART-EMAP [9], PROBART [10], FasArt and FasBack [11]. Besides, Gaussian ARTMAP (GA) [12] and Distributed ARTMAP (DA) [13] also address the noisy data to some extent. (2) FA employs axis-parallel hyperrectangles as category template to summarize data into groups, which is an issue of representation error when the sample distribution is rather complex. Several variants of the FA attempt to utilize nonrectangular categories such as hyperspheres in Hypersphere ARTMAP (HA) [14], hyperellipsoids in Ellipsoid ARTMAP (EA) [15] and Gaussian category choice function in GA. However, none of these inner categories can approximate the borders with flexibility. The correspondence between the geometries of data set and the internal categories is still an important factor in the performance of ART networks.

The recent Polytope ARTMAP (PTAM) [16] is an alternative architecture to the FA in the sense that it uses a different geometry for category formation and representation. The utility of general, irregular polytope geometry for the internal categories makes PTAM not specially suit to any particular geometry and provides less dependence on data set geometry than the other ART networks. Hence, PTAM is clearly better than the best rectangular and circular ART networks on a data set with irregular geometry. Experimental results [16] on a complete collection of 2-D data sets show that PTAM achieves lower error than the leading ART networks, with a similar number of categories, while achieving higher error in the presence of noise or statistical overlapping between the classes. PTAM approximates the borders among the output predictions by strictly following the information contained in the training set, both in the category expansion and adjustment steps. Flexible category representation without statistical information in the category expansion and adjustment steps does minimize the representation error while making the approximation error out of control in the presence of noise or prediction overlap.

However, another recently proposed architecture using the Bayesian framework, called Bayesian ARTMAP (BA) [17], proves superior performance, with respect to noise and sensitivity to statistical overlapping. BA's most interesting and appealing property is that the accuracy of the BA is very close, and sometimes even identical to the Bayes' bound of accuracy for all the experiments on 1-D synthetic data [17]. In this paper we propose Bayesian Polytope ARTMAP (BPTAM) that modifies some of the characteristics of the PTAM algorithm by a similar methodology employed by the BA, which attempts to combine the computational advantages of both PTAM and BA. BPTAM incorporating both simplex categories and Gaussian categories produces comparable category representation error while ameliorating the approximation error by replacing the deterministic rules with statistical learning and inference. During training, the simplex categories expand only towards the input pattern without category overlap, while the Gaussian categories grow or shrink by limiting their hypervolumes. During testing, which kind of inner geometry category should be adopted to compete and resonate depends on the calculation of choice function (CCF) for the current input pattern. The BPTAM is evaluated here in comparison with BA and PTAM using different experiments on synthetic data. After optimizing BPTAM, it shows better applicability to data sets with noise, statistical overlapping as well as irregular geometry.



978-1-4244-2794-9/09/$25.00 ©2009 IEEE

The remaining of the paper is organized as follows. Section II presents a detailed description of BPTAM and explains how the methodology used in BA can be naturally extended to BPTAM. Section III shows some preliminary experimental results, which clearly demonstrate the applicability of our approach. We are going to discuss these experimental results in section IV and finally summarize the main conclusions in section V.

## II. BAYESIAN POLYTOPE ARTMAP

The categories in PTAM are irregular polytopes delimited by hyperplanes which compose a piecewise linear approximation to the borders among output predictions. The Polytope vertices are weight vectors selected among previous training patterns in order to define these borders. PTAM categories have no predefined geometry, although they are internally managed as a set of adjacent simplexes, and each CCF is a combination of the simplexes choice functions. Since the simplex covers the smallest volume defined by its vertices, the polytope category can expand only towards the input pattern **I**, and not in other directions, by adding a new simplex between them or replacing a vertice with **I**. PTAM replaces the vigilance test in the ART networks by Overlap Test (OT), which itself is not enough to avoid category overlap. Prediction Test (PT) offers a complement way to adjust previous wrong expansions. If the active category C passes OT and PT after expansion, the input pattern is assigned to the category. If no category passes either test, PTAM creates a new one. Whether it can be a polytope category or a single vector depends on the OT and the number of weight vectors from the same prediction. In the presence of noise or prediction overlap, the input patterns belonging to different categories are mixed and they break simplexes in the category adjustment steps, which leads to be the creation of noisy single-vector categories. This operation suffers the performance of PTAM due to the invalidation of summarizing data into groups. In this scenario, PTAM attempts to memorize inherent noise, which is the phenomenon of category proliferation.

The overall aim of the BPTAM research program is to reduce the sensitivity to noise or statistical overlapping and develop a modification to PTAM which can show better generalization capability on various kinds of data sets with noise, statistical overlapping as well as irregular geometry. In this section, we will present some important aspects of BPTAM such as its major components, the way it describes categories, its operation and the way it performs learning.

### *A. Overview of Bayesian PTAM Architecture*

Due to PTAM's design that follows the operating principles of FA, modifications made in the past to FA in order to improve some of its shortcomings can be readily and easily applied to PTAM. Considering that we will never know the exact borders among output predictions when the classifier performs on-line learning, irregular geometry which is designed to approximate these borders only by OT and PT cannot cover the regions efficiently. Under some geometrical circumstances, the regions which are not covered by simplexes may be blank or be surrounded by some trivial simplexes. [18] proposed Distributed PTAM to address this problem using distributed learning. However, the proposed BPTAM here employs the main stages of PTAM while replacing the deterministic rules with statistical learning and inference. There are several aspects of the new architecture that should be pointed out.

First, two kinds of inner geometry categories, simplex categories and multidimensional Gaussian categories, provide better representation of the complex distributed data, while all the other ART networks using only one shape of categories. Flexible representation of simplex categories without overlap, adding Gaussian categories which permit overlapping with other Gaussian categories and simplex categories, gives no opportunity to the creation of noisy single-vectors. Second, the huge difference between irregular polytope and multidimentional Gaussian makes the calculation of choice function (CCF) of each category suit to its shape respectively. Specifically, the CCF of Gaussian categories using Bayes' decision theory accounts not only the distance of a category to a pattern but also to the dominance of category with respect to other categories through the category prior probability. Each input pattern should go though the CCF of polytope categories first and then the CCF of Gaussian categories. Third, in the testing phase, the input pattern which locates in the inner part of any irregular simplexes employs the deterministic associations learned by those simplex categories, otherwise using the probabilistic associations learned previously by those Gaussian categories. These operations ameliorate the overall performance of BPTAM and make this classifier more adaptable to different data sets.

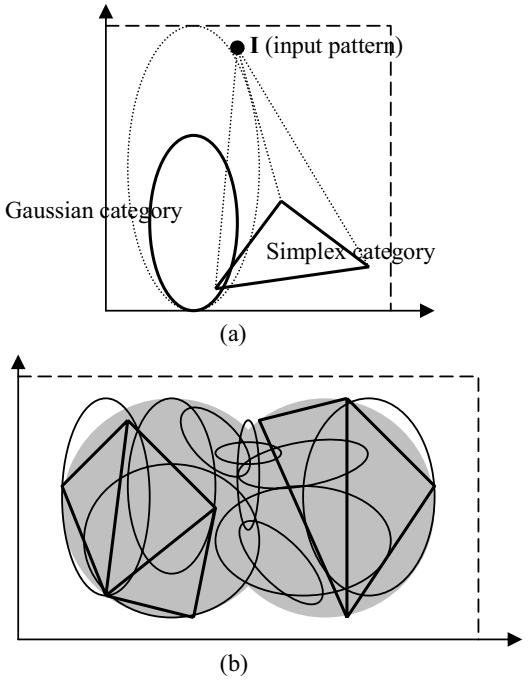


Figure 1. Two Examples of geometrical categories used in BPTAM. (a) An input pattern **I** learned by a simplex category and a Gaussian category simultaneously. (b) The region covered by simplex categories and Gaussian categories when confronting a data set with two overlapped Gaussian distributions.

*B. Bayesian PTAM Categories*

Fig. 1 gives two examples of geometrical categories employed by BPTAM, which illustrates the order in which categories are searched. During the training phase of BPTAM, learning is accomplished by creating new categories or by expanding already existing ones. As shown in Fig. 1(a), the current input pattern **I** is going to be encoded into both Gaussian categories and simplex categories. The simplex only expands towards the input pattern, not in other directions, while the Gaussian category will expand enough to include this pattern in its representation region. Each input pattern can at least be encompassed into one Gaussian category even when both OT and PT fail and no simplex expansion conducts. In this case, no opportunity is left for the creation of noise single-vectors which leads to category proliferation. Fig. 1(b) illustrates the region covered by both simplex categories and Gaussian categories in two overlapped Gaussian distributions. Simplex categories can only appear in the purer regions with the same output predictions, while the Gaussian categories occupy almost everywhere. In addition, the more complex and overlapped the region is, the more trivial Gaussian categories appear. These Gaussian categories instead of single-vector categories are probabilistically associated with class, which can provide better generalization.

*C. Operation of Bayesian PTAM*

In this subsection, we describe three main aspects of the training and testing phase in BPTAM. The interested reader may refer to [16] and [17] for more specification of the same operations as PTAM and BA. Herein, we only demonstrate how the methodology used in BA extends PTAM to address its deficiencies in the Bayesian framework.

1) **Calculation of CCF**. There are two types of CCF in BPTAM for simplex category and Gaussian category respectively. The simplex category's CCF $T_i(\mathbf{I})$ is defined as the maximum of choice function $T_{ij}(\mathbf{I})$ of its simplexes $S_{ij}$, j = 1. . . . . $N_{is}$ :

$$T_i(\mathrm{I}) = \max\{T_{ij}(\mathrm{I})\}, \qquad i = 1.....N_c \tag{1}$$

Where,

$$T_{ij}(\mathrm{I}) = \begin{cases} 1 & g_{ijk}(\mathrm{I}) > 0 \qquad k = 1.....n+1 \\ e^{-d(\mathrm{I},S_{ij})/\gamma}, & otherwise \end{cases} \tag{2}$$

$T_{ij}(\mathbf{I})$ is defined in such a way that $g_{ijk}(\mathbf{I}) > 0$ if I falls inside the simplex $S_{ij}$: in this case, $T_i(\mathbf{I}) = T_{ij}(\mathbf{I}) = 1$ because **I** falls inside the category representation region (CRR) of $C_i$. Otherwise, $0 < T_{ij}(\mathbf{I}) < 1$, and it decreases with the distance $d(\mathbf{I}, S_{ij})$ between the input pattern and the simplex. During training, the calculation of Gaussian category's CCF is right after that of simplex category. The *a posteriori* probability of the jth Gaussian category to represent the M-dimensional pattern is computed by

$$\mathrm{M}_j = P(w_j \mid x) = \frac{p(x \mid w_j)\,P(w_j)}{\sum_{l=1}^{N_{cat}} p(x \mid w_j)\,P(w_j)} \tag{3}$$

Where $N_{cat}$ is the number of categories, $P(w_j)$ is the estimated prior probability of the jth Gaussian category and $p(\mathbf{x}|w_j)$ is the likelihood of $w_j$ with respect to **x**.

$$p(x \mid w_j) = \frac{1}{(2\pi)^{M/2} \mid \sum_j \mid^{1/2}} \exp\{-0.5(x-\mu)^T \sum_j^{-1} (x-\mu)\} \tag{4}$$

Where $\mu_j$ and $\Sigma_j$ are the estimated mean and covariance matrix of jth Gaussian category. The chosen Gaussian category J is the one with the maximum *a posteriori* probability. Each input pattern should choose a Gaussian category, but the simplex category is not necessary if the OT and PT fail during training.

2) **Learning.** If the chosen simplex category passes both the OT and PT which are the same as in PTAM, then it has the right prediction and it expands towards the input pattern without overlap. Whether the simplex category expands or not, single-vectors still have no chance to appear. This property is derived from the learning of those Gaussian categories. Similarly to the BA, the vigilance test ensuring that the chosen Gaussian category is limited in size makes BPTAM not vigilance-free any more. If the Jth Gaussian category's hypervolume matches the maximal hypervolume, then the Gaussian category parameters are updated in the presentation of this current input pattern. No matter whether the simplex category goes to resonance with this input pattern or not, the chosen Gaussian category will be adjusted by the following equations.

$$\mu_{J,new} = \frac{N_J}{N_J+1}\mu_{J,old} + \frac{1}{N_J+1}x \tag{5}$$

$$\sum_{J,new} = \frac{N_j}{N_J+1}\sum_{J,old} + \frac{1}{N_J+1}(x-\mu_{J,new})(x-\mu_{J,new})^T * I \tag{6}$$

Where $N_J$ is the number of patterns that have been clustered by the Jth Gaussian category before introducing the current pattern and *I* is the identity matrix.

3) **Inference**. In the training phase, the BPTAM performs two kinds of geometry category learning, which leads to the creation of double associations between categories and classes. In the testing phase, the calculation of simplex categories' CCF $T_i(\mathrm{I})$ determines which association can be selected to inference. This combined inference methodology helps BPTAM deal with different kinds of data sets. If the borders among the output predictions are rather complex to approximate, the simplex categories and the corresponding association operate. The Gaussian categories and the probabilistic association can perform better confronting more naturally distributed data.

***If*** $T_i(\mathbf{I})$==1

Chose the association learned by the simplex categories.

***Else***

Chose the association learned by the Gaussian categories.

The class chosen for a test pattern **x** is

$$c_I = \arg\max_i P(c_i \mid x) \tag{7}$$

***End***

Where,

$$P(c_i \mid x) = \frac{P(c_i, x)}{p(x)} = \frac{\sum_{j=1}^{N_{cat}} P(c_i, w_j, x)}{p(x)} = \frac{\sum_{j=1}^{N_{cat}} P(c_i \mid w_j) p(x \mid w_j) P(w_j)}{\sum_{k=1}^{C} \sum_{j=1}^{N_{cat}} P(c_i \mid w_j) p(x \mid w_j) P(w_j)} \quad (8)$$

Where $P(c_i|w_j)$, $P(w_j)$,and $p(x|w_j)$ have been defined in [17].

## III. PRELIMINARY EXPERIMENTATION

In this section we present some illustrative results to demonstrate the capability of BPTAM as a classifying machine. Towards that end we compare BPTAM with PTAM and BA in the respects of the optimization process, learning curves, classification accuracy and category proliferation.

### A. Experimental Setup

Fig. 2 shows three kinds of data sets used in our experimental work, which demonstrates the sensitivity to the noise, statistical overlapping and the irregular geometry. Data sets CIS_noise showed in Fig. 2(a) are generated by adding Gaussian noise to the CIS data sets. The noise level is used, given by the standard deviation of the Gaussian distribution, with value of 0.05 in the experiment. The data set $4G_3$ illustrated in Fig. 2(b) is used to evaluate classification algorithms when predictions overlap, with four output predictions, given by Gaussian probability distributions with standard deviation 0.05. The overlap extent is determined by the distance between its centers and the center of unit square. Data Set with Irregular Geometry, not specially suited to circular or rectangular category geometries, has two predictions associated to the input patterns falling inside and outside of the curve in Fig. 2(c).

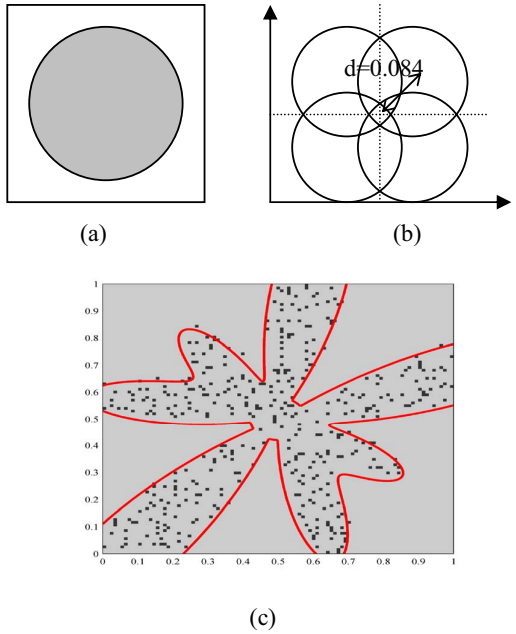


Figure 2. (a) CIS_noise data, (b) $4G_3$ data. (c) Data Set with Irregular Geometry used in the experimental work.

### B. Optimization

The maximal Gaussian category hypervolume parameters in both BPTAM and BA were optimized to the previously mentioned classification tasks, over the range $10^{-8} < S_{max} < 10^{2}$. The optimal parameter values for both BA and BPTAM were determined based on the comparatively high classification accuracy and low number of categories created on a validation set of 1000 patterns independent of the training and test sets. Fig. 3 shows the BA and BPTAM validation accuracies as well as the number of categories recorded for increasing maximal Gaussian category hypervolume values, with 100 training patterns for both CIS_noise and $4G_3$ data set while 500 training patterns for Data Set with Irregular Geometry. This figure can be roughly divided into three different regions. The first region, denoted as (1), represents overfitting by the BA and BPTAM with the highest validation accuracy as well as the highest number of categories created for $4G_3$ and Data Set with Irregular Geometry, while achieving relatively high validation accuracy for CIS_noise data set. On the opposite, the region (3), representing underfitting by the BA and BPTAM, yields almost the lowest validation accuracies for all the data sets used in our experimental work except for the Data Set with Irregular Geometry. The validation accuracy for Data Set with Irregular Geometry is nearly the same as the highest accuracy achieved by the BA in region (1). That is, the performance of BPTAM for the third classification task is more stable than the BA with varying maximal Gaussian category hypervolume, which shows the applicability of BPTAM to the complex distributed data. Region (2) is where the maximal Gaussian category hypervolume parameter should be picked up. Based on the relatively high accuracy as well as the relatively low number of categories, we choose $S_{max}$ to be $10^{-5}$. In addition, the validation accuracy is enhanced almost in every region for all the data sets, while the categories created in BPTAM are increased. This is due to the utility of two kinds of inner categories for one input pattern.

### C. Learning Curves

The purpose of this experiment is to investigate the learning curves of the BA, PTAM and BPTAM by measuring their test accuracy and number of categories evaluated on different sizes of training patterns. For CIS_noise data set, we employ 1000 patterns as the test set for the three kinds of data sets. Using the optimal maximal Gaussian category hypervolume, all of the three classifiers are trained on sets of increasing sizes (from 100 to 1000 patterns in increments of 100 patterns). In Fig. 4, the BPTAM has superior test accuracy to PTAM on CIS_noise and $4G_3$ while superior to BA on data set with irregular geometry. Although the categories created are more than that of the BA and PTAM, we still emphasize that the BPTAM can enhance the model classification accuracy with less sensitivity to noise, statistical overlapping and complex distribution with irregular geometry, which shows its applicability to different kinds of data sets.

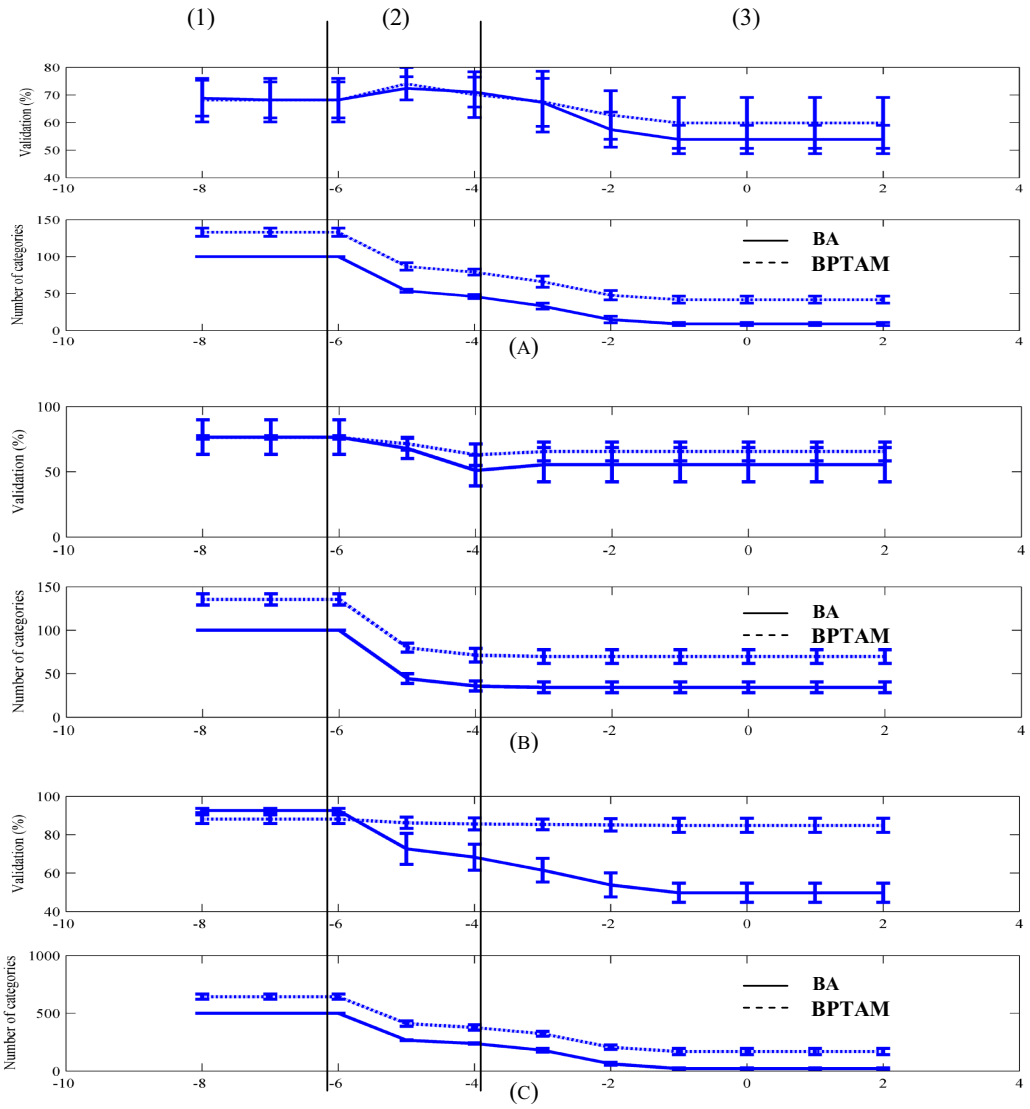


Figure 3. The comparison of optimization process in BA and BPTAM on (a) CIS_noise Data Set, (b) $4G_3$ Data Set, (c) Data Set with Irregular Geometry, respectively.

## IV. Discussion

The proposed BPTAM preserves both the advantages of BA and PTAM in the Bayesian framework. This is due to the fact that the utility of both simplex categories and Gaussian categories provides more flexibility in the category representation and more associations between categories and classes in a probabilistic fashion. Literature [17] reports that the BA manifests high accuracy, stable learning curves, and a small, constant number of categories even with a small sample size. We modified PTAM in order to enhance its classification accuracy while wishing to address the category proliferation problem. However, the model of BPTAM is as much sensitive to the sample size as the other classifiers except for BA. Fig. 3 illustrates that BPTAM ameliorate the validation accuracy in the price of more categories which is composed of simplex categories and Gaussian categories. Fig. 4 also demonstrates how the category growth with the sample size performs. The good results of the BA with respect to the category proliferation are possible when classification tasks are evaluated on data sets with 1-D Gaussian and non-Gaussian densities, which have been demonstrated in [17]. Our experiments extends the research with 2-D CIS data sets with Gaussian noise, 2_D $4G_3$ data set with four Gaussians overlapping and 2-D complex distributed data with irregular geometry. Although small $S_{max}$ leads to the proliferation of Gaussian categories to some extent, BPTAM enhance the stability of classification accuracy.

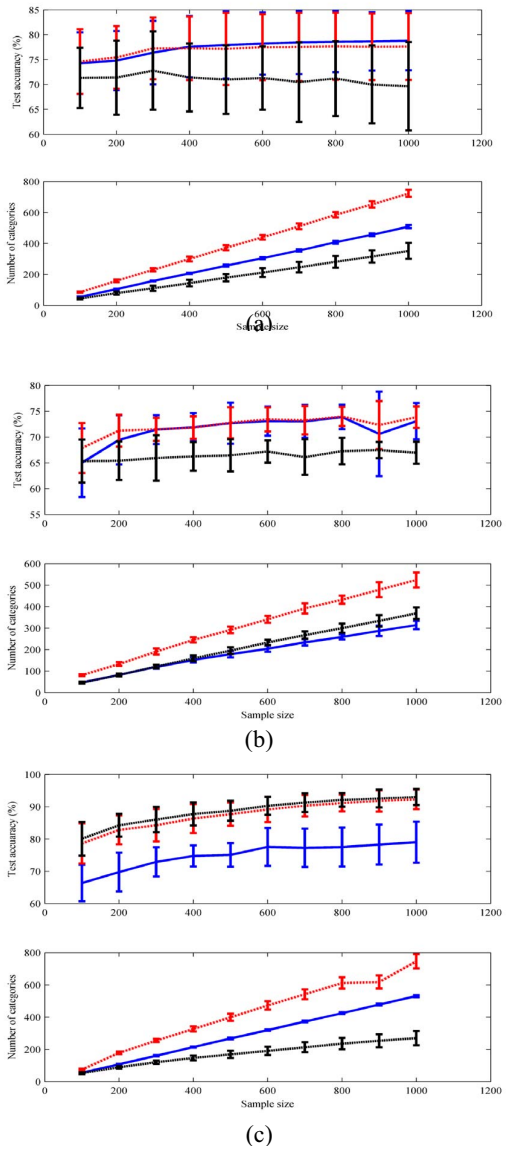


Figure 4. The comparison of learning curves of BA(blue), PTAM(black) and BPTAM(red) on (a) CIS_noise Data Set, (b) $4G_3$ Data Set, (c) Data Set with Irregular Geometry, respectively.

## V. CONCLUSION

In this paper we introduced Bayesian Polytope ARTMAP (BPTAM) as a variant of Polytope ARTMAP (PTAM) that can potentially reduce its generalization error. This is derived from the utility of two kinds of inner geometry categories, the employment of Bayes' decision theory for learning and inference as well as the probabilistic associations between categories and classes. Flexible category representation with statistical information in the category expansion and adjustment steps does minimize the representation error while simultaneously making the approximation error under control in the presence of noise or prediction overlap. We also present some preliminary but illustrative experimental results that show the potential of BPTAM as a classifier when confronting data sets with noise, statistical overlapping and irregular geometry.

## REFERENCES


[1] S.Grossberg, "Adaptive Pattern Classification and Universal Recoding, I: Parallel Development and coding of neural feature Detectors,"Biological Cybernetics, vol. 23, pp. 121–134, 1976.

[2] G.A.Carpenter and S.Grossberg, "Massively parallel architecture for a self-organizing neural pattern recognition machine," Computer Vision, Graphics, and Image Processing, vol. 37, pp. 54–115, 1987.

[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 Networks, vol. 4, pp. 565–588, 1991.

[4] G.A.Carpenter, S.Grossberg, N.Markuson, J.H.Reynolds, and D.B.Rosen, "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps," IEEE Trans. On Neural Networks, vol. 3(5), pp.698–713, 1992.

[5] S.J.Verzi and G.L.Heileman, "Generalization Performance of ART-Based Networks in Structural Risk Minimization Framework," In Proc. Int. Joint Conf. Neural Netw., 2002, pp.2644-2649.

[6] S.J.Verzi, G.L.Heileman, M.Georgiopoulos, and M.J.Healy, "Boosting the performance of ARTMAP," In Proc. Int. Joint Conf. Neural Netw., 1998, pp.369-401.

[7] Eduardo Gómez-Sánchez, Tannis A. Dimitradis, José Manuel Cano-Lzquierdo, and Juan López-Coronado, "μARTMAP: Use of Mutual Information for Category Reduction in Fuzzy ARTMAP," IEEE Trans. On Neural Networks, vol. 13(1), pp. 58–69, 2002.

[8] S.J.Verzi, G.L.Heileman, M.Georgiopoulos, and M.J.Healy, "Hierarchical ARTMAP," In Proc. Int. Joint Conf. Neural Netw., 2000, pp.41-46.

[9] 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(4), pp.805-818, 1995.

[10] S.Marriott, and R.F.Harrison, "A modified Fuzzy ARTMAP architecture for the approximation of noisy mappings," Neural Networks, vol. 8(4), pp.619-641, 1995.

[11] José Manuel Cano Izquierdo, Yannis A. Dimitriadis, Eduardo Gómez Sánchez, Juan López Coronado, "Learning from noisy information in FasArt and FasBack neuron-fuzzy systems," Neural Networks, vol.14, pp.407-425, 2001.

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

[13] G. A Carpenter, B.L. Milenova, B.W. Noeske, "Distributed ARTMAP: a neural network for fast distributed supervised learning," Neural Netw., pp.793-813, 1998.

[14] G.CAnagnostopoulos, and M.Georgiopoulos, "Hypersphere ART and ART-MAP for unsupervised and supervised incremental classification," In Proc. Int. Joint Conf. Neural Netw., 2000, pp.59-64.

[15] G. C Anagnostopoulos, and M. Georgiopoulos, "Ellipsoid ART and Ellipsoid ARTMAP for incremental clustering and classification," In Proc. Int. Joint Conf. Neural Netw., 2001, pp.1221-1226.

[16] Dinani Gomes Amorim, Manuel Fernández Delgado, and Senén Barro Ameneiro, "Polytope ARTMAP: Pattern Classification Without Vigilance Based on General Geometry Categories," IEEE Trans. Neural Netw., vol.18(5), pp.1306-1325, 2007.

[17] Boaz Vigdor, and Boaz Lerner, "The Bayesian ARTMAP", IEEE Trans. Neural Netw., vol.18(6), pp.1628-1644, 2007.

[18] Leonardo Liao, and Yongqiang Wu, "Distributed Polytope ARTMAP: A vigilance-free ART network for distributed supervised learning," In Proc. Int. Joint Conf. Computational Sciences and Optimization, 2009, in press.