Hypersphere ART and ARTMAP for Unsupervised and Supervised, Incremental Learning

Georgios C. Anagnostopoulos,
Michael Georgiopoulos
School of Electrical Engineering & Computer Science
University of Central Florida, Orlando, Florida, U.S.A.
gca@bruce.engr.ucf.edu, michaelg@mail.ucf.edu

Abstract

A novel adaptive resonance theory (ART) neural network architecture is being proposed. The new model, called Hypersphere ART (H-ART) is based on the same principals like Fuzzy-ART does and, thus, inherits most of its qualities for unsupervised learning. Among these properties is fast, stable, incremental learning on the training set and good generalization on the testing set. While H-ART is intended for clustering tasks, its extension, H-ARTMAP is playing the role of Fuzzy-ARTMAP's counterpart for the supervised learning of real-valued, multi-dimensional mappings. Also in this paper, some experimental results are presented involving the comparison of H-ARTMAP and Fuzzy-ARTMAP in simple, illustrative classification problems. The results are indicating comparable performances in error rate but also a good potential for substantial superiority of H-ARTMAP in terms of nodes (categories) utilized. The latter effect can be attributed to H-ART's more efficient internal knowledge representation.

I. Introduction

Fuzzy-ART and *Fuzzy-ARTMAP* were first introduced in [1] and [2] as a self-organizing, parallel clustering and a pattern classification machine respectively; they both originated from the pioneering work presented in [3] on *adaptive resonance theory* (ART) and extend the family of ART architectures by being capable of performing learning tasks with real-valued patterns. Properties of these architectures have been investigated in [1], [4], [5] and [6]. Also, numerous applications based on the same networks have been reported in the literature. For the length of this paper it is assumed that the reader is already familiar with the basics of these networks.

In brief, Fuzzy ART performs a clustering of its input feature space into *categories* during its training phase. The category representation process takes place in the F_2 layer of its *attentional subsystem.* This layer consists of *committed nodes* sensitized to each category. The internal representation of categories is materialized by means of *templates*, one associated to each node. Geometrically, each template in Fuzzy-ART can be represented as a *hyperrectangle* in the input feature space. All input patterns contained in such a hyperrectangle are said to be encoded by the corresponding template. The number and size of these templates and, hence, the magnitude of input data summarization is determined during the training phase by Fuzzy-ART's parameters: the *vigilance parameter* ρ, the *choice parameter* α and the *uncommitted node choice parameter* M_u. The formation of templates within the network is accomplished incrementally using off-line or on-line training. All the patterns of the training set are presented one by one to the network. Each pattern will either select an already existing category or initiate the creation of a new one, if no existing category can satisfactorily represent it. In the latter case, we say that the pattern chooses the *uncommitted node.* The selection process is based on the *vigilance test*, which assesses the proximity of the pattern to a category in terms of city block (L_1 norm) distance, and the size of the its related hyperrectangle. The smaller and closer a category's hyperrectangle is in relation to the pattern, the more the potential for this category to be selected by the pattern. Once a category is selected, its associated committed node resonates. During resonance, update of its template occurs only if the pattern is not enclosed by the node's hyperrectangle; otherwise, no learning occurs. In particular, when off-line training is being performed, eventually all patterns will select a committed node without causing a template update. This condition signifies the conclusion of the training phase. Once trained, the performance phase follows similar rules to associate a previously "unseen" pattern to one of the network's clusters or to the uncommitted node, which would imply that the pattern most likely does not originate from the same population as the training patterns do. Up to this point, the sketching of Fuzzy-ART's behavior has been attempted; some information about Fuzzy-ARTMAP will be provided in a later section.

The rest of the paper is organized as follows. In the subsequent section the motivation behind H-ART and H-ARTMAP is being exposed along with their governing learning rules. Furthermore, a comparison is made between the geometrical representations of Fuzzy-ART and H-ART in a 2-dimensional setting to demonstrate their

0-7695-0619-4/00 $10.00 © 2000 IEEE

analogies. Section III presents illustrative classification tasks that compare the performance of Fuzzy-ARTMAP and H-ARTMAP. A simple, 2-dimensional classification problem and an additional benchmark task example have been chosen to accommodate a comparison of the two architectures. The primary focus of the comparison lies in the classification error rate and the magnitude of node proliferation. Finally, Section IV summarizes the material presented and the experience accumulated during the investigation of the experimental results.

II. Description of H-ART and H-ARTMAP

As was pointed out in [7], in many clustering or classification problems the way input patterns are distributed in the feature space does not lend itself for accurate summarization using the ranges of values defined by Fuzzy-ART's hyperrectangles. This effect is more intense in the presence of noise and leads to category proliferation in order to compensate for the lack of representation efficiency. In other words, excessive amounts of nodes may have to be employed in the F_2 layer in some cases to achieve a reasonable accuracy when predicting categories. It seems that circles, spheres and, in general, hyperspheres are for many problems a more natural and efficient geometrical representation of boundaries between clusters in the feature space. Based on these facts as a motivating point, *Gaussian-ART* and *Gaussian-ARTMAP* were introduced in [7], which addressed this issue successfully using hyperellipsoids for the geometric representation. However, one of the major disadvantages of these two architectures is the lack of finite, stable learning. With Fuzzy-ART and Fuzzy-ARTMAP fast learning completes in a finite number of list presentations. At the end of the training phase the network will predict the correct category for each one of the input patterns used for training. This is not the case with the two former architectures. Overtraining under fast learning causes the category representations to gradually reduce in size causing the encoding of patterns, that might have been already learned by the network, to deteriorate.

The focus of the work presented in this paper was to capture an ART neural scheme combining the virtues of both approaches. The end product is *Hypersphere-ART* (H-ART) and its corresponding ARTMAP (H-ARTMAP) architecture. H-ART, as its name suggests, utilizes *hyperspheres* for the geometric representation of clusters. From that aspect, a hypersphere network should be able in general to employ fewer nodes for its learning tasks, as opposed to its fuzzy counterpart. Even though hyperspheres are potentially less efficient in comparison to hyperellipsoids with respect to category shape description, they indeed allow for a template update law that supports finite, stable fast learning.

As we noted earlier, H-ART encodes and thus describes clusters of patterns in the input feature space as hyperspheres. Since hyperspheres embedded in an $\mathbb{R}^M$ feature space are completely determined by their radius R and their centroid $\mathbf{m}$, a template (bottom-up weight) of H-ART is of the form $\mathbf{w} = (R, \mathbf{m}) \in \mathbb{R}^{M+1}$. It is worth noting that H-ART is using only M+1 memory elements per node in contrast to Fuzzy-ART that needs $2M$ per node to characterize each of its hyperrectangles. This follows from the fact that Fuzzy-ART's templates are of the form $\mathbf{w} = (\mathbf{u}, \mathbf{v}^c) \in \mathbb{R}^{2M}$, where $\mathbf{u}$ and $\mathbf{v}$ are the points defining Fuzzy-ART's hyperrectangle. The notation $\mathbf{v}^c$ denotes the vector $\mathbf{1}$-$\mathbf{v}$, where $\mathbf{1} \in \mathbb{R}^M$ is the all-ones vector. At this point it becomes apparent that another difference between the two architectures is that H-ART lacks a F_0 layer; complementary encoding of input patterns is unnecessary under the new ART scheme. On the other hand though, a form of pre-processing still has to be performed on the input patterns to compute a lower bound on H-ART's $\bar{R}$ parameter, which is not shared by Fuzzy-ART and will be mentioned later in the text. However, both H-ART and Fuzzy-ART have exactly the same operation characteristics in common, that is, their training phase and performance phase are of identical fashion. Due to their functional similarity, H-ART inherits all of Fuzzy-ART's parameters, namely the vigilance ρ, the committed node choice parameter α and the uncommitted node choice parameter $\bar{R}_u$. Moreover, the major role of these parameters is carried over to H-ART. The vigilance controls the maximum size of templates that can be formed and in conjunction with $\bar{R}_u$ (denoted M_u for Fuzzy-ART) regulates the number of categories created during learning. Finally, the choice parameter α mainly influences the competition among nodes for the encoding of a pattern to be learned by the network.

When an input pattern $\mathbf{x} \in \mathbb{R}^M$ is presented to the F_1 layer, a bottom-up input $T(\mathbf{x}|\mathbf{w}_j)$, also called *category choice function*, is generated for each node j of the F_2 layer. For H-ART the Weber Law form of the choice function for the j^{th} committed node is

$$T(\mathbf{x} \mid \mathbf{w}_j) = \frac{\bar{R} - \max\{R_j, \| \mathbf{x} - \mathbf{m}_j \|_2\}}{\bar{R} - R_j + \alpha} \tag{1}$$

In the previous equation, $\|.\|_2$ denotes the L_2 vector norm. The choice parameter a takes values in the interval $(0, \infty)$. The H-ART specific $\bar{R}$ parameter is called *radial extend* of the network and takes values in the range $[R_{max}, \infty)$. Its primary role is to control the maximum size that categories can potentially reach during the training phase. The lower bound R_{max} is the *feature space radius* defined as

$$R_{\max} = \frac{1}{2} \max_{i,j} \| \mathbf{x}_i - \mathbf{x}_j \|_2 \tag{2}$$

The pre-processing required estimating R_{max} is of $o(P^2/2)$ complexity, where P is the number of available patterns. The estimate is used to choose a practical value for the radial extend parameter before commencing training. In addition, we define the choice function of the uncommitted node as

$$T_u = \frac{\bar{R}}{\bar{R}_u + \alpha} \tag{3}$$

It will be assumed that $\bar{R}_u = 2\bar{R}$. The node J eventually chosen by the pattern (the resonating node) will be the one with the highest choice function value that satisfies the *vigilance criterion*

$$\rho(\mathbf{x} \mid \mathbf{w}_J) \geq \rho \tag{4}$$

where $\rho(\mathbf{x}|\mathbf{w}_j)$ is the *category match function* for the j^{th} node and is defined as

$$\rho(\mathbf{x} \mid \mathbf{w}_j) = 1 - \frac{\max\{R_j, \| \mathbf{x} - \mathbf{m}_j \|_2\}}{\bar{R}} \tag{5}$$

As in the Fuzzy-ART case, the vigilance parameter ρ takes its values in the interval $[0, 1]$. If the resonating node J is a committed one, then its template will be updated in the manner described below

$$\begin{aligned} R_{J,new} &= R_{J,old} + \frac{\gamma}{2}\left(\max\{R_{J,old}, \| \mathbf{x} - \mathbf{m}_{J,old} \|_2\} - R_{J,old}\right) \\ \mathbf{m}_{J,new} &= \mathbf{m}_{J,old} + \frac{\gamma}{2}\left(1 - \frac{\min\{R_{J,old}, \| \mathbf{x} - \mathbf{m}_{J,old} \|_2\}}{\| \mathbf{x} - \mathbf{m}_{J,old} \|_2}\right)(\mathbf{x} - \mathbf{m}_{J,old}) \end{aligned} \tag{6}$$

Here, $\gamma \in (0, 1]$ stands for the learning rate characterizing the network's training mode. For $\gamma = 1$ we have fast learning and any other value implies slow learning. If it happens that none of the committed nodes satisfy the vigilance constraint and the uncommitted node is chosen, then a new category is created. The node's template will be initialized as follows

$$\begin{aligned} R_J &= 0 \\ \mathbf{m}_J &= \mathbf{x} \end{aligned} \tag{7}$$

Although not apparent at first glance, Eq. 6 implements a template update method for H-ART, which resembles the update mechanism of Fuzzy-ART. To be more specific, Eq. 6 allow a node to learn a pattern without forgetting patterns already encoded by the same node. This fact can be demonstrated with the simple example shown in Fig. 1. The left side of Fig. 1 depicts a geometrical illustration of an already committed node resonating upon presentation of pattern **x** in a 2-dimensional feature space setting. It is further assumed that this node with template $\mathbf{w}_{old}$ satisfies the vigilance criterion and that fast learning is in effect ($\gamma = 1$). Its geometrical representation, the circle C_{old}, will expand enough to become C_{new} by adjusting its centroid $\mathbf{m}_{old}$ and radius R_{old} to include **x**. The adjustments obey the rules set in Eq. 6 and are a function of the euclidian distance $dis(\mathbf{x}, C_{old})$ between the pattern to be encoded and the original circle C_{old}. For comparison reasons, the right side of Fig. 1 illustrates the same update procedure for a Fuzzy-ART template $\mathbf{w}_{old}$ under the same assumptions. Instead of a circle, in the general case the geometrical

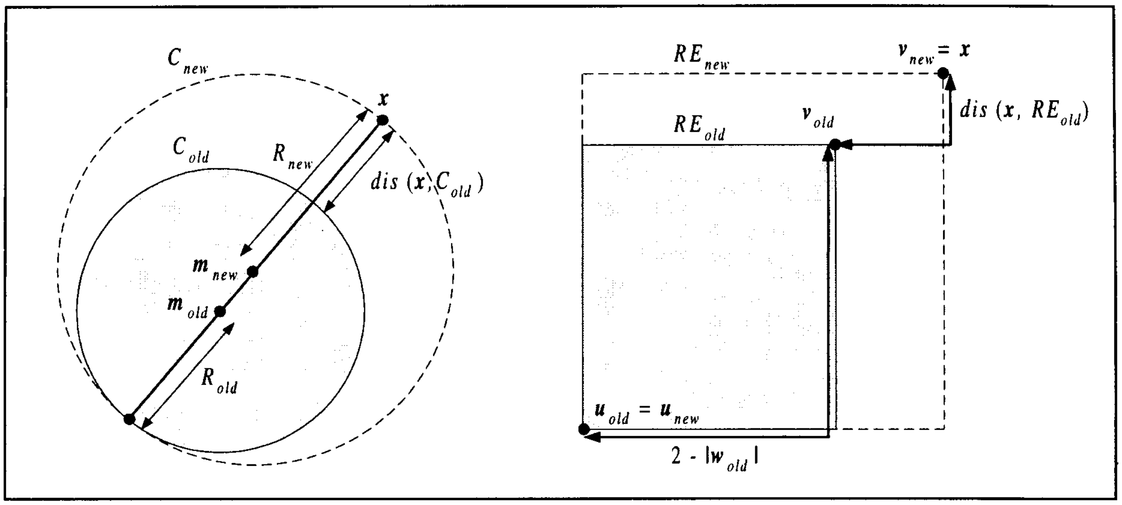

Figure 1

representation of a Fuzzy-ART template would be a rectangle. Upon resonance, the template with representation RE_{old} will expand enough to include the pattern and form RE_{new}. Again, the modifications are a function of the L_l distance dis(x, RE_{old}) between the pattern and the pre-resonance rectangle RE_{old}. Notice, that when ever a pattern belongs to the closure (interior or boundary) of a template representation, then dis(x, C_{old}) = 0 for H-ART and dis(x, RE_{old}) = 0 for Fuzzy-ART. Under these circumstances no template update is taking place for Fuzzy-ART. According to Eq 6, the same fact holds for H-ART as well, since in this case $R_{old} \geq \|x-m_{old}\|_2$. This learning behavior is of vital importance for both architectures, since it aids in equipping the networks with the *self-stabilization* property [1].

The learning scheme presented so far via Eq. 1 through Eq. 7 attributes H-ART with the same properties as Fuzzy-ART: fast, stable, incremental clustering of the input space. Its performance phase is identical to the training scheme described so far with the exception that no updates are implemented to the network's templates. However, instead of choosing the category with the maximum value of bottom-up inputs in a *winner-take-all* (WTA) fashion, alternative approaches can be considered including the *Q-max rule* [8] and others. In essence, WTA is a special case of the Q-max rule for $Q = 1$.

As mentioned before, H-ARTMAP is the counterpart of Fuzzy-ARTMAP and, thus, consists of two ART modules operating under the H-ART learning and performance rules presented. Each one of them is associated to a feature space; ART$_a$ clusters the input patterns and ART$_b$ the output patterns. It has to be noted that the values of the learning/network parameters (the vigilance, the choice parameters and the radial extend) of each ART module might be chosen independently of each other. The modules interface with each other via an inter-ART module that maps categories of the input feature space with categories in the output feature space. In addition, the training phase behavior of ART$_a$ is augmented with *match tracking*. Without going into extensive detail, match tracking attempts to rectify erroneous association of an ART$_a$ category with an ART$_b$ category by suitably increasing the vigilance in the ART$_a$ module. Therefore, in the case ARTMAP networks the initial value of ρ is called *baseline vigilance value.* As a general comment, all similarities between H-ART and Fuzzy-ART induce corresponding similarities to their ARTMAP counterparts. When an ARTMAP architecture is used for pattern classification tasks, ART$_b$ nodes encode the labels of each class to be learned by using a vigilance value equal to one. Therefore, ART$_b$ templates and parameters are of no practical interest [6]. In the sequel, ARTMAP parameters will imply parameters in ART$_a$. In order to demonstrate the similarities but also some potential differences between H-ARTMAP and Fuzzy-ARTMAP, the next section exhibits some limited comparison results and attempts to reflect the experience gathered in the process.

III. Experimental Results

Two experiments are being considered in this section. The first one is the "circle in a square" problem [7] and deals with artificially generated data from between two non-overlapping classes. More specifically, the classifier has to discriminate between the points inside (the "inside"-points) and outside of a circle (the "outside"-points) embedded inside the unit-area square. The second experiment conducted was to classify the PIMA Indian Diabetes Database (PID) [9]. Both the experiments were performed using some common settings and rules for both architectures. First, fast learning ($\gamma = 1$) was used and training was performed to perfection, that is, until all training patterns were correctly classified. Secondly, the performance phases were conducted using the WTA rule and $\rho = 0$ to force a choice among the given class labels. Finally, for both phases of operation it was assumed that $M_u = 2M$ and $\overline{R}_u = 2\overline{R}$.

For the "circle in a square" classification task, 200 uniformly distributed training patterns were generated. Approximately equal number of patterns was sampled from each class. Training was performed by experimenting with a variety of parameter values. For each set of values 100 different orders of pattern presentation were considered. The evaluation of Fuzzy-ARTMAP and H-ARTMAP was accomplished be using a testing set of 1000 patterns to assess their classification error rates. The latter quantity is defined as being the fraction of testing patterns misclassified. Figure 2 reflects part of the results produced via a percent error rate vs. number of categories formed. An ideal classifier would achieve minimal classification error by employing minimal number of nodes. The plot indicates the performance of Fuzzy-ARTMAP with points and the one of H-ARTMAP with crosses. Evidently, H-ARTMAP achieves the lowest error with the lowest number of categories. Using only two nodes it exhibits 99.99% correct classification, since it misclassified only one test pattern. A single node represented the interior of the circle and a second one the exterior. Their corresponding circles were centered in the vicinity of point (0.5, 0.5) and the one encoding the "outside"-points enclosed the other one. Although the "inside"-patterns were enclosed by both circles, they would correctly chose the "inside"-node, since it was the one with the smallest size (radius). H-ARTMAP's optimal performance was reached with $\overline{R} = 2$, $\rho = 0.4$ and $a = 10^{-4}$. On the other hand, Fuzzy-ARTMAP exhibited a comparable performance of 96.6% utilizing 116 nodes. This result was achieved using $\rho = 0.1$ and $a = 10$.

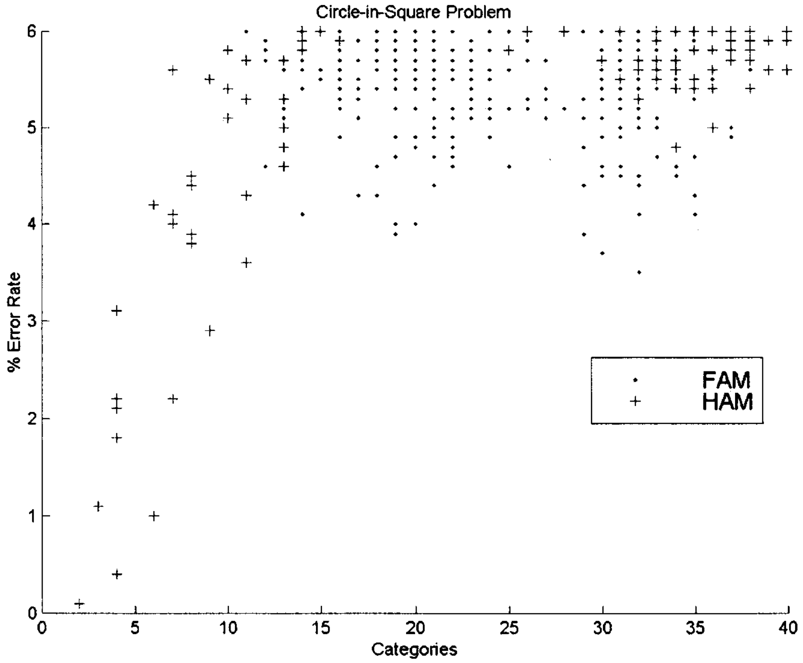

Figure 2

For the PID problem the available data was bisected into 576 training and 192 testing samples. The input feature space is of dimensionality 8 and the patterns belong to two classes. During experimentation only the case with $\rho = 0$ was investigated. H-ARTMAP yielded optimal classification of 73.43% and employed 115 nodes. Fuzzy-ARTMAP's best instance exhibited less correct classifications (only 66.30% with 62 nodes) using small values of the choice parameter. An interesting fact is that although H-ARTMAP in this example creates more categories (almost double the amount than Fuzzy-ARTMAP, both of them utilize about the same amount of memory. This is because H-ARTMAP uses 9 memory elements per node in comparison to the 16 that Fuzzy-ARTMAP requires.

There are some general comments that can be made at this point. First, H-ARTMAP usually requires higher vigilance values to reach the same performance as Fuzzy-ARTMAP. The influence of H-ARTMAP's choice parameter diminishes for high values of the radial extend, which is also implied by Eq. 6. Moreover, selecting a good value for $\bar{R}$ is dependent on the problem at hand. The experience gained from the experimentation on the aforementioned problems suggests values much higher than the input feature space radius. This seems to allow for higher data summarization ability. Another interesting fact about H-ARTMAP is that, when ρ, α and $\bar{R}$ are chosen such that a single node is generated for each training pattern, then the network's response is equivalent to the *k-nearest neighbor* (KNN) classification method [10] with $k = 1$ for WTA or $k = Q$ for a Q-max performance rule. Finally, even when H-ARTMAP creates more nodes than Fuzzy-ARTMAP for a specific level of error rate, in the vast majority of cases it will utilize, in general, less memory.

IV. Conclusions

A new ART neural architecture for clustering was presented, namely H-ART, and its extension H-ARTMAP for classification tasks. A comparison was drawn between these architectures and the original Fuzzy-ART and Fuzzy-ARTMAP. Since H-ART is based on the underlying concepts of Fuzzy-ART, it inherits most of its properties, the most important one being the ability for incremental, finite and stable learning. Due to its more efficient internal representation of knowledge, H-ART suggests a potential for achieving similar performance to Fuzzy-ART with less categories created and less memory needed. The same properties carry over to H-ARTMAP for classification problems. Only limited insight was provided by two illustrative classification examples. Additional experiments, though, need to be conducted to validate some of the above claims regarding H-ART and H-ARTMAP. The authors are currently investigating different approaches to improve H-ART and H-ARTMAP, as well as hybrid hyperrectangle/hypersphere network schemes for achieving highest data compression without compromising classification rates.

V. References

[1] Carpenter, G.A., Grossberg, S., & Rosen, D.B. (1991) "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system", Neural Networks, 4, 759-771.

[2] Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J., & Rosen, D.B. (1992) "Fuzzy-ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps", IEEE Transactions on Neural Networks, 3, 698-713.

[3] Grossberg, S. (1976) "Adaptive pattern recognition and universal recoding II: Feedback, expectation, olfaction, and illusions", Biological Cybernetics, 23, 187-202.

[4] Huang, J., Georgiopoulos, M., & Heileman, G.L. (1995) "Fuzzy ART properties", Neural Networks, 8(2), 203-213.

[5] Georgiopoulos, M., Huang, J., & Heileman, G.L. (1994) "Properties of learning in ARTMAP", Neural Networks, 7, 495-506.

[6] Georgiopoulos, M., Fernlund, H., Bebis, G., & Heileman, G.L. (1996) "Order of search in Fuzzy ART and Fuzzy ARTMAP: Effect of the Choice Parameter", Neural Networks, 9(9), 1541-1559.

[7] Williamson, J.R. (1996) "Gaussian ARTMAP: A Neural Network for Fast Incremental Learning of Noisy Multidimensional Maps", Neural Networks, 9(5), 881-897.

[8] Carpenter, G.A., Markuzon, N. (1998) "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases", Neural Networks, 11, 323-336.

[9] Murphy, P.M., & Aha, D.W. (1992) "UCI repository of machine learning databases", Department of Information & Computer Science, University of California, Irvine, CA.

[10] Duda, R.O., & Hart, P.E. (1973) "Pattern classification and scene analysis", New York, Wiley.