ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 495-500

Convergence properties of a fuzzy ARTMAP network

Răzvan Andonie [1] and Lucian Sasu [2]

[1]Computer Science Department, Central Washington University, USA

[2]Computer Science Department, Transylvania University of Braşov, Romania

Abstract. FAMR (Fuzzy ARTMAP with Relevance factor) is a FAM (Fuzzy ARTMAP) neural network used for classification, probability estimation [3], [2], and function approximation [4]. FAMR uses a relevance factor assigned to each sample pair, proportional to the importance of that pair during the learning phase. Due to its incremental learning capability, FAMR can efficiently process large data sets and is an appropriate tool for data mining applications. We present new theoretical results characterizing the stochastic convergence of FAMR.

1 Introduction

An *incremental learning* algorithm can be defined by the following characteristics [12]: *i)* it is able to learn additional information from new data; *ii)* it does not require access to the original data, used to train the existing system; *iii)* it preserves previously acquired knowledge; and *iv)* it is able to accommodate new data categories that may be introduced with new data.

When designing and implementing data mining applications for large data sets, we face processing time and memory space problems. In this case, incremental learning is a very attractive feature. In the context of supervised training, incremental learning means learning each input-output sample pair, without keeping it for subsequent processing. Very few algorithms perfectly fit into this description of incremental learning. The FAM family of neural networks [5] is the best known example. The FAM model has been incorporated in the MIT Lincoln Lab system for data mining of geospatial images because of its computational capabilities for incremental learning, fast stable learning, and visualization [11].

Many pattern recognition applications require an estimate of the *posterior* probability $P(C|\mathbf{a})$, where C is a class index and $\mathbf{a}$ is an input pattern. This task also allows classification because one can select the class C with the maximum conditional probability. Another classical application of neural networks

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 495-500

is the prediction (approximation) of functions that are known only at a certain number of points.

FAMR is a FAM-based neural network used for classification, posterior probability estimation [2], [3], and function approximation [4]. It is a generalization of another FAM architecture: PROBART [10]. FAMR uses a relevance factor assigned to each sample pair, proportional to the importance of that pair during the learning phase. This adds more flexibility to the training phase, allowing ranking of sample pairs according to the confidence we have in the information source. The training sequence may include sample pairs from sources with different levels of noise.

We analyze how the relevance factors influence the convergence of the learning process in a FAMR. Without answering completely this question, we describe here several convergence properties of FAMR. In Section 2 we review the basic notations and the learning algorithm used in FAMR. Section 3 analyzes the stochastic convergence of FAMR, and Section 4 contains conclusions and open problems.

2 A description of FAMR

A FAM consists of a pair of fuzzy ART modules, ART_a and ART_b, connected by a an inter-ART module called Mapfield. For a presentation of FAM, we suggest [9]. The details of the FAMR architecture can be found in [3].

The FAMR learning paradigm is based on the following stochastic approximation procedure. Let us consider a sequence of independent experiments according to the finite probability distribution $P(a_1), \ldots, P(a_n)$, where $P(a_i) \geq 0$ is the probability of outcome a_i, $\sum_{i=1}^{n} P(a_i) = 1$. These *objective probabilities* are not known and will be estimated at each step based on the previous observations. A criterion for a qualitative differentiation of the experiments is represented by the relevance associated to each experiment. The *relevance* q_t is a real positive finite number directly proportional to the importance of the experiment considered at step t ($t = 1, 2, \ldots$). For example, one could assign the relevance factor according to the confidence in the training data source, when several sources are used. The relevance can be also related to the closeness of the decision boundaries.

The following estimation procedure makes use of both the results and the relevances of the present and previous experiments.

The *subjective probability* of outcome a_i ($i = 1, \ldots, n$) at step t ($t = 1, 2, \ldots$) is given by:

$$w_t(a_i) = w_{t-1}(a_i) + A_t \left(\delta_t(a_i) - w_{t-1}(a_i)\right) \tag{1}$$

where: if at step t we get outcome a_j, $\delta_t(a_j) = 1$, and $\delta_t(a_i) = 0$ for $j \neq i$; $w_0(a_i) \geq 0$ is the initial subjective probability, $\sum_{i=1}^{n} w_0(a_i) = 1$; $q_0 \geq 0$ is the initial relevance, $Q_t = \sum_{s=0}^{t} q_s$, and $A_t = q_t/Q_t$ ($A_t = 0$ for $Q_t = 0$).

Theorem 1. *$w_t(a_i) \xrightarrow{t} P(a_i)$ in probability iff $Q_t \xrightarrow{t} \infty$.*

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 495-500

Proof. The proof can be found in [1]. □

Let $w_t^{(n)}(a_i)$ be the subjective probabilities at step t ($t = 1, 2, \ldots$), for n possible outcomes. What is happening if at some step we get a new outcome, a_{n+1}? Assuming we have $w_0^{(n)}(a_i) = 1/n$ ($i = 1, \ldots, n$), the new subjective probabilities $w_t^{(n+1)}(a_i)$ for $n+1$ possible outcomes can be obtained by the following relations:

$$\begin{cases} w_t^{(n+1)}(a_{n+1}) = q_0/(n+1)Q_t \\ w_t^{(n+1)}(a_i) = w_t^{(n)}(a_i) - w_t^{(n+1)}(a_{n+1})/n,\ i = 1, 2, \ldots, n \end{cases} \tag{2}$$

Relations (2) will be used in the dynamic allocation of ART_b categories (*Step 2* in Algorithm 1.)

Mapfield weight w_{jk}^{ab} can be considered an estimate of the posterior probability $P(k|j)$. This enables us to use formula (1) to update the weights w_{jk}^{ab}:

$$w_{Jk}^{ab(new)} = \begin{cases} w_{jk}^{ab(old)} & \text{if } j \neq J \\ w_{JK}^{ab(old)} + A_t(1 - w_{JK}^{ab(old)}) & \\ w_{Jk}^{ab(old)}(1 - A_t) & \text{if } k \neq K \end{cases} \tag{3}$$

Let $\mathbf{Q}$ be the vector $[Q_1 \ldots Q_{N_a}]$. N_a and N_b are the number of categories in ART_a, respectively ART_a, initialized with 0. The training pair formed by input vector a and output vector b can be learned by the FAMR Mapfield algorithm given in Algorithm 1.

Based on Theorem 1, can we say that w_{jk}^{ab} a good estimate of $P(I_{\mathbf{b}}|I_{\mathbf{a}})$, where $I_{\mathbf{a}}$ and $I_{\mathbf{b}}$ are intervals based around input pattern $\mathbf{a}$, respectively output pattern $\mathbf{b}$? Feedback via match tracking alters this estimation (see [10]). One way to avoid this problem is to eliminate match tracking.

Eliminating match tracking is not always convenient, because it controls category proliferation in ART_a. Meanwhile, it is difficult to say something about this probability approximation in the presence of match tracking, since in this case w_{jk}^{ab} is not necessarily a good estimate of the posterior probability with respect to the already processed data. However, in our experiments, match tracking has not significantly altered probability estimation.

3 Theoretical results

We aim to analyze more carefully relation (1) and the FAMR learning algorithm.

By imposing restrictions to q_t, Theorem 1 can be strengthened, and a convergence rate can be computed. The restrictions imposed to q_t are natural: an observer who intends to learn objective probabilities from examples has to have sufficient confidence in the results of the experiences.

Theorem 2. *For $0 \leq q_0 < \infty$, $0 < a \leq q_t < \infty$ ($t = 1, 2, \ldots$), $w_t(a_i)$ converges in the mean square to $P(a_i)$.*

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 495-500

Step 1. Accept $(\mathbf{a}, \mathbf{b})$ with relevance factor q.
Step 2. If necessary, create a new category K in ART_b:
$N_b = N_b + 1$, $K = N_b$
if $N_b > 1$ **then**
$w_{jK}^{ab} = \frac{q_0}{N_b Q_j}$ for $j = 1, \ldots, N_a$ {append new component to $\mathbf{w}_j^{ab}$}
$w_{jk}^{ab} = w_{jk}^{ab} - \frac{w_{jK}^{ab}}{N_b - 1}$ for $k = 1, \ldots, K-1$; $j = 1, \ldots N_a$ {normalize}
endif
Step 3. If necessary, create category J in ART_a:
$N_a = N_a + 1$, $J = N_a$
$Q_J = q_0$ {append new component to $\mathbf{Q}$}
$w_{Jk}^{ab} = 1/N_b$ for $k = 1, \ldots, N_b$ {append new line to $\mathbf{w}^{ab}$}
Step 4. J, K are winners or new added nodes
if $N_b\, w_{JK}^{ab} \geq \rho_{ab}$ **then**
{learn in Mapfield}
$Q_J = Q_J + q$
$w_{JK}^{ab} = w_{JK}^{ab} + \frac{q}{Q_J}(1 - w_{JK}^{ab})$
$w_{Jk}^{ab} = w_{Jk}^{ab}\left(1 - \frac{q}{Q_J}\right)$ for $k = 1, \ldots, N_b$, $k \neq K$
else
perform match tracking and restart from step 3
endif

Algorithm 1: **One iteration in the FAMR Mapfield algorithm.**

Proof. For any $t \geq 1$, we obtain from (1):

$$E\left(w_t(a_i) - P(a_i)\right)^2 = \frac{q_0^2\left(w_0(a_i) - P(a_i)\right)^2}{Q_t^2} + \frac{1}{Q_t^2} E\left(\sum_{s=1}^{t} q_s(\delta_s - P(a_i))\right)^2 + + \frac{2q_0\left(w_0(a_i) - P(a_i)\right)}{Q_t^2} E\left(\sum_{s=1}^{t} q_s(\delta_s - P(a_i))\right) \quad (4)$$

Since $\{\delta_s - P(a_i)\}_{s \geq 1}$ are zero biased independent random variables, we have:

$$E\left(w_t(a_i) - P(a_i)\right)^2 = \frac{q_0^2\left(w_0(a_i) - P(a_i)\right)^2}{Q_t^2} + \frac{P(a_i)(1 - P(a_i))\sum_{s=1}^{t} q_s^2}{Q_t^2} \quad (5)$$

Applying Stoltz's Lemma, it results: $E\left(w_t(a_i) - P(a_i)\right)^2 \xrightarrow{t} 0$. □

Theorem 3. *For $q_0 \in [0, b]$, $q_t \in [a, b]$ $(t = 1, 2, \ldots)$, where a and b are any two real numbers $0 < a \leq b < \infty$, we have: $w_t(a_i) \xrightarrow{t} P(a_i)$ with probability one.*

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 495-500

Proof. Our proof is based on the Stochastic Approximation Theorem [7]. We have obtained an alternative proof using Kolmogorov's criterion [6]. □

Let $S_t = \sum_{s=1}^{t} q_s^2$. We can compute an upper limit for the convergence rate $w_t(a_i) - P(a_i)$.

Theorem 4. *If conditions in Theorem 3 are true then, for any real $\epsilon > 0$, the following inequality holds almost surely, except for a finite set of terms:*

$$\begin{aligned} |w_t(a_i) \quad - \quad P(a_i)| \le |w_0(a_i) - P(a_i)|\frac{q_0}{Q_t} + \\ + \quad (1+\epsilon)\frac{\sqrt{2P(a_i)(1-P(a_i))S_t \log\log\left[P(a_i)(1-P(a_i))S_t\right]}}{Q_t} \end{aligned} \quad (6)$$

Proof. We use a LIL–type theorem for martingales (see [8]), for the martingale $w_t(a_i) - P(a_i)$. The powerful martingale theory allows us to bypass the restriction for random variables to be independently and identically distributed, required in the limit theorems of classical probability theory. □

We can state now our main result:

Theorem 5. *If match tracking is not used then, for each ART_a category j $(j = 1, \ldots, N_a)$ and each ART_b category k $(k = 1, \ldots, N_b)$, we have:*

1. *If $0 \le q_0 < \infty$, $0 < a \le q_t < \infty$ $(t = 1, 2, \ldots)$, then $\mathbf{w}_{JK}^{ab}$ converges in the mean square to $P(K|J)$;*
2. *If $q_0 \in [0, b]$, $q_t \in [a, b]$ $(t = 1, 2, \ldots)$, where a and b are any two real numbers $0 < a \le b < \infty$, then $\mathbf{w}_{JK}^{ab} \xrightarrow{t} P(K|J)$ both with probability one and in the mean square, and the convergence rate has almost surely the upper limit (6), except for a finite set of terms.*

4 Conclusions and problems to be investigated

The FAMR algorithm expands the range of FAM applications by allowing to assign a relevance factor to each training pair. The FAMR learning process in the Mapfield module is based on stochastic approximation and is influenced by the relevance factors of the training pairs. Our theoretical results characterize the convergence and the convergence rate of the approximation. We have at least two open problems:

1. Assuming that we have a set of vector pairs with fixed relevances, how do we select from it an optimal training set?
2. Given a training set, how do we optimally assign relevances to the training pairs?

"Optimal" in both cases refers to providing the fastest FAMR learning.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 495-500

References

[1] R. Andonie. A converse H-theorem for inductive processes. *Computers and Artificial Intelligence*, 9:159–167, 1990.

[2] R. Andonie and Sasu L. A fuzzy ARTMAP probability estimator with relevance factor. In M. Verleysen, editor, *Proceedings of the 11th European Symposium on Artificial Neural Networks (ESANN2003)*, pages 367–372, Bruges, Belgium, April 23-25 2003.

[3] R. Andonie, L. Sasu, and V. Beiu. Fuzzy ARTMAP with relevance factor. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2003)*, pages 1975–1980, Portland, Oregon, July 20-24 2003.

[4] R. Andonie, L. Sasu, and V. Beiu. A modified fuzzy ARTMAP architecture for incremental learning function approximation. In O. Castillo, editor, *Neural Networks and Computational Intelligence*, volume Anaheim, California: ACTA Press, Proceedings of the IASTED International Conference on Neural Networks and Computational Intelligence (NCI 2003), pages 124–129, Cancun, Mexico, May 19-21 2003.

[5] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, 3(5):698–713, 1992.

[6] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. John Willey & Sons Inc., 1970.

[7] R. M. Golden. *Mathematical Methods for Neural Network Analysis and Design*. MIT Press, 1996.

[8] P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Applications*. Academic Press, 1980.

[9] C. P. Lim and R. F. Harrison. ART-based autonomous learning systems: Part I - Architectures and Algorithms. In L. C. Jain, B. Lazzerini, and U. Halici, editors, *Innovations in ART Neural Networks*. Springer, 2000.

[10] S. Marriott and R. F. Harrison. A modified fuzzy ARTMAP architecture for the approximation of noisy mappings. *Neural Networks*, 8(4):619–641, 1995.

[11] O. Parsons and G. A. Carpenter. ARTMAP neural network for information fusion and data mining: map production and target recognition methodologies. *Neural Networks*, 16:1075–1089, 2003.

[12] R. Polikar, L. Udpa, S. S. Udpa, and V. Honovar. Learn++: An incremental learning algorithm for supervised neural networks. *IEEE Transactions on Systems, Man, and Cybernetics–PartC*, 31(4):497–508, 2001.