Overtraining in Fuzzy ARTMAP: Myth or Reality?

Michael Georgiopoulos[1], Anna Koufakou[2], Georgios C. Anagnostopoulos[3], Takis Kasparis[4]

University of Central Florida
School of Electrical Engineering & Computer Science
4000 Central Florida Blvd.
Engineering Building, Room 407
Orlando, Florida, United States 32816
Orlando, Florida, US
[1]michaelg@mail.ucf.edu, [2]akoufakou@yahoo.com, [3]anagnostop@email.com,
[4]kasparis@pegasus.cc.ucf.edu

Abstract

In this paper we are examining the issue of overtraining in Fuzzy ARTMAP. Over-training in Fuzzy ARTMAP manifests itself in two different ways: (a) it degrades the generalization performance of Fuzzy ARTMAP as training progresses, and (b) it creates unnecessarily large Fuzzy ARTMAP neural network architectures. In this work we are demonstrating that overtraining happens in Fuzzy ARTMAP and we propose an old remedy for its cure: cross-validation. In our experiments we compare the performance of Fuzzy ARTMAP that is trained (i) until the completion of training, (ii) for one epoch, and (iii) until its performance on a validation set is maximized. The experiments were performed on artificial and real databases. The conclusion derived from these experiments is that cross-validation is a useful procedure in Fuzzy ARTMAP, because it produces smaller Fuzzy ARTMAP architectures with improved generalization performance. The trade-off is that cross-validation introduces additional computational complexity in the training phase of Fuzzy ARTMAP.

1 Introduction

Fuzzy ARTMAP has been introduced in the neural network literature by Carpenter, et al., 1992, and since then it has been established as one of the premier neural network architectures in solving classification problems. In solving classification problems Fuzzy ARTMAP has the capability of establishing arbitrary mappings between clusters of an input space of arbitrary dimensionality and clusters of an output space of arbitrary dimensionality. At times, in doing so it creates very large neural network architectures. As a result, a number of researchers have tried to address this problem with various degrees of success (e.g., see Williamson, 1996, Vertzi, et al., 1998, and Gomez Sanchez, et al, 2000). In Vertzi, et al, 1998, the authors discussed the issue of overtraining in Fuzzy ARTMAP. This issue is most apparent when the classes of the classification problem that Fuzzy ARTMAP tries to solve exhibit significant overlap and results in the creation of large Fuzzy ARTMAP neural network architectures. In this paper we address the same problem, the problem of overtraining in Fuzzy ARTMAP. Overtraining in Fuzzy ARTMAP manifests itself in two different ways. It may decrease the generalization performance of the network or it may increase the size of the Fuzzy ARTMAP architecture (without necessarily improving its generalization), or both. To address the problem of overtraining in Fuzzy ARTMAP we propose the usage of cross-validation techniques. Cross validation is a well-respected procedure in the statistical literature that allows you to determine when overtraining occurs. To avoid some of the issues that plague cross-validation approaches (e.g., the issue of small dataset) we focus our attention here only on databases that have sufficient number of datapoints. This way, we can split the data into a training, validation and test set that are representative of the distribution that the data follow. The type of Fuzzy ARTMAP networks investigated are (a) a Fuzzy ARTMAP network that is trained until completion, (b) a Fuzzy ARTMAP network that is trained for one epoch, and (c) a Fuzzy ARTMAP network that is trained to the point where its performance on the validation set is maximized. Our careful examination of the literature did not identify any references where Fuzzy ARTMAP training is stopped early through a cross-validatory procedure. As we have mentioned earlier this is the topic that this paper addresses. The details of the Fuzzy ARTMAP neural network architecture are included in Carpenter, et. al, 1992.

0-7803-7044-9/01/$10.00 ©2001 IEEE

2 Cross-Validation

Estimating the accuracy of a classifier induced by supervised learning methods, such as Fuzzy ARTMAP, is an important issue. One of the reasons for its importance is that it gives us some guidance on how good the future predictive accuracy of the classifier is. Another, equally important reason, is that it gives us a way of choosing the "best" classifier model amongst a set of classifier models.

Cross-validation is a statistical technique that allows us to estimate the accuracy of a classifier model. Kohavi, 1995, discusses two prominent cross-validation procedures. The first one referred to as the holdout method. We split the set S of available data into a training set S_{tr} and a validation set S_v. The classifier is designed using the data in the training set S_{tr} and its accuracy is estimated by evaluating its performance on the validation set S_v. That is, the holdout estimated accuracy is defined as

$$PCC_v = 100\frac{1}{NV}\sum_{(I_i,O_i)\in S_v}\delta(y_i,O_i) \quad (1)$$

where PCC_v denotes the percentage of correct classification of the classifier over the validation set S_v, NV are the number of datapoints in validation set S_v, the I_i and O_i designate the i-th input and desired output pair in S_v, y_i is the actual response of the classifier when it is excited by the input I_i, and $\delta(x,y)=1$ if $x=y$, while $\delta(x,y)=0$ if $x\neq y$.

Obviously the holdout estimate is a random number that depends on the division of the available data in S into a training set S_{tr} and a validation set S_v. Often the holdout method is repeated k times and the estimated accuracy PCC_v is produced by averaging the estimated accuracies of the k runs.

The second method for cross-validation is referred to as k-fold cross-validation. In this procedure the available data S is split into k mutually exclusive subsets, designated as S^1, S^2, ..., S^k of approximately equal size. The classifier is trained and tested (validated) k times. Each time m, $m\in\{1, 2, ..., k\}$, it is trained on $S\backslash S^m$ and tested on S^m. The cross-validation estimate is defined as the number of correct classifications divided by the number of data points in the set S. That is,

$$PCC_v = 100\frac{1}{NV}\sum_{m=1}^{k}\sum_{(I_i,O_i)\in S^m}\delta(y_i,O_i) \quad (2)$$

where PCC_v is the percentage of correct classification on the validation set (which in this case happens to be the entire set of available data), NV is the number of elements in S_v (which happens to be the same as S), (I_i, Oi) represents a generic input/desired output pair in S^m, and y_i is the actual output of the classifier, designed with data in $S\backslash S^m$, and excited with the input I_i from the set S_m. Once more, $\delta(x,y)=1$ if $x=y$, while $\delta(x,y)=0$ if $x\neq y$.

Obviously the cross-validation estimate in equation (2) is a random number that depends on the division into folds. Complete cross-validation is the average of the above estimates over all the possible folds of NT training data into k folds of approximately equal size. This is too expensive though, except in the case of 1-fold cross-validation, with NT relatively small. As Kohavi states repeating cross-validation multiple times using different splits into folds provides a better estimate at the expense of additional computational cost. In stratified cross-validation, the folds are stratified so that they contain approximately the same proportions of labels as the original set.

In this paper we use stratified cross-validation to stop training of Fuzzy ARTMAP at a point where its performance on the validation set is maximized. To produce the estimate of the Fuzzy ARTMAP performance we used the holdout cross-validation technique. Since we are focusing on datasets with large number of data samples we do not have to worry about making inefficient use of the available data. Furthermore, since we deal with large databases we did not use k-fold cross-validation to avoid increased computational costs.

3 Experiments - Results - Observations

We have conducted experiments with artificial databases to demonstrate the potential of cross-validation in Fuzzy ARTMAP. The artificial databases consist of Gaussian data that are of dimensionality 2 or 5 or 10. They belong to either 2 different classes or 3 different classes. The degree of overlap of data that belong to different classes is either low, or medium, or high. The Gaussian data generated are independent in different dimensions and their means and variances are chosen appropriately so that they can justify the characterization of low, medium, or high overlap.

For example, let us assume that we have a collection of Gaussianly distributed data, of dimensionality 2, that belong to 2 different classes. We decided to use 5,000 datapoints per class to train Fuzzy ARTMAP (this set is S_{tr}), 5,000 different datapoints per class to cross-validate Fuzzy ARTMAP (this set is S_v), and 5,000 different datapoints per class to test the performance of the trained Fuzzy ARTMAP (this set is S_{tes}). We trained Fuzzy ARTMAP in three different modes:

1. Mode 1: Train Fuzzy ARTMAP with the training data until completion (i.e., until Fuzzy ARTMAP's misclassification rate on the training data is 0%). Evaluate the performance of the trained Fuzzy

ARTMAP on the test data (S_{tes}). This performance is denoted by PCC^{c}_{tes}.

2. Mode 2: Train Fuzzy ARTMAP for one complete epoch (an epoch of training corresponds to one presentation of all input/output pairs of the training set through Fuzzy ARTMAP). Evaluate the performance of the trained Fuzzy ARTMAP on the test data (set S_{tes}). This performance is denoted by PCC^{1EP}_{tes}.

3. Mode 3: Train Fuzzy ARTMAP for one complete epoch but check its performance on the validation set (set S_v) every 100 iterations of training (an iteration of training corresponds to one input/output training pair presentation to Fuzzy ARTMAP). At the end of the one epoch of training we identify the iteration number at which the trained Fuzzy ARTMAP has exhibited the maximum performance on the validation set. We denote this performance as PCC_v. The weights of the Fuzzy ARTMAP that exhibited the maximum performance on the validation set are retained. These weights are then used to evaluate Fuzzy ARTMAP's performance on the test set (set S_{tes}). We denote this performance by PCC_{tes}.

For all the aforementioned three modes of training we also retained the information about the number of nodes that the trained Fuzzy ARTMAP has created. We denote the number of these nodes as N^{c}_{a}, N^{1EP}_{a} and N_a, for modes 1, 2 and 3 of training, respectively. For the artificial databases Mode 3 cross-validation was performed only for the first epoch of training, due to the fact that cross-validation is a computationally expensive procedure. We observed that for the artificial databases performing cross-validation only for the 1st epoch of training was enough, since we were able to produce a small Fuzzy ARTMAP architecture with a good generalization performance.

Our experimental results with the artificial databases are illustrated in Table 1. In Table 1 we depict the results in 8 different columns. Column 1, designated, as Overlap defines the degree of overlap between the data belonging to different classes. The second column of Table 1 depicts the number of classes in our dataset; as we have mentioned before we have experimented with data belonging to 2 or 3 distinct classes. The third column in Table 1 shows the dimensionality of the input patterns. To discuss the rest of the columns of Table 1, let us focus on one of the rows of Table 1, the boldfaced entry of the medium overlap category corresponding to data of dimensionality 10, belonging to 3 classes. The results reported in columns 4 through 8 of the boldfaced entry of the medium overlap category are extracted by averaging the results over 25 experiments. These experiments were constructed by taking 5 different sets of training/validation/test data and for each such set of data we trained Fuzzy ARTMAP with 5 distinct orders of training data presentations. For future reference we refer to these 5 different sets of data as S^{m}_{tr}, S^{m}_{v}, and S^{m}_{tes}, for $1 \le m \le 5$. For each one of these sets we refer to the 5 orders of training data presentation by $or(m)$, where $or(m)$ takes the values 1, 2, 3, 4, 5 to designate the five different orders of presentation for each one of the 5 training data sets. The entry of the fourth column of the boldfaced row in the medium overlap category corresponds to $\overline{\overline{PCC^{c}_{tes}}}$. The entry of the fifth column of the boldfaced row in the medium overlap category corresponds to $\overline{\overline{PCC_{tes}}} - \overline{\overline{PCC^{c}_{tes}}}$. The quantities $\overline{\overline{PCC_{tes}}}$ and $\overline{\overline{PCC^{c}_{tes}}}$ are defined as follows:

$$\overline{\overline{PCC_{tes}}} = \frac{1}{25}\sum_{m=1}^{5}\sum_{or(m)=1}^{5} PCC_{tes}(m, or(m)) \tag{3}$$

$$\overline{\overline{PCC^{c}_{tes}}} = \frac{1}{25}\sum_{m=1}^{5}\sum_{or(m)=1}^{5} PCC^{c}_{tes}(m, or(m)) \tag{4}$$

where $PCC_{tes}(m,or(m))$ is the performance of Fuzzy ARTMAP on the test data S^{m}_{tes}, trained under mode 3, with training data S^{m}_{tr} presented to it in the order $or(m)$, while $PCC^{c}_{tes}(m,or(m))$ is the performance of Fuzzy ARTMAP on the test data S^{m}_{tes}, trained under mode 1, with training data S^{m}_{tr} presented to it in the order $or(m)$.

Note that the entries of the fourth column of Table 1, which correspond to the average percentage of correct classification for Mode 1 Fuzzy ARTMAP (complete training scenario), are a quantitative verification that we are dealing with a low, medium or high overlap. The $\overline{\overline{PCC^{c}_{tes}}}$ value for the low overlap is in the high 90's range, the medium overlap is in the low to mid-80's range and the high overlap is in the 60's to 70's range. The entry of the sixth column of the boldfaced row in the medium overlap category corresponds to $\overline{\overline{PCC_{tes}}} - \overline{\overline{PCC^{1EP}_{tes}}}$, which is the average difference in the percentage of correct classification between the Mode 3 and Mode 1 trained Fuzzy ARTMAPs. The seventh column, designated as $\overline{\overline{CR^{c}}}$, corresponds to the average ratio of the number of nodes created by the Mode 1 trained Fuzzy ARTMAP and the number of nodes created by the Mode 3 trained Fuzzy ARTMAP. This ratio is referred to as compression ratio complete (CR^c), to remind us how much Mode 3 trained Fuzzy ARTMAP compresses the information compared to Mode 1 trained Fuzzy ARTMAP (which is trained to completion). The eighth column, designated as $\overline{\overline{CR^{1EP}}}$, corresponds to the average ratio of the number of nodes created by the Mode 2 trained Fuzzy ARTMAP and the number of nodes created by the Mode 3 trained Fuzzy ARTMAP. This ratio is referred to as compression ratio one epoch (CR^{1EP}), to remind us how much Mode 3 trained Fuzzy ARTMAP compresses the information compared to Mode 2 trained Fuzzy ARTMAP (which is trained for one epoch). The definitions of the quantities $\overline{\overline{PCC_{tes}^{1EP}}}$, $\overline{\overline{CR^c}}$, and $\overline{\overline{CR^{1EP}}}$ are similar with the definitions of the quantities $\overline{\overline{PCC_{tes}}}$ and $\overline{\overline{PCC_{tes}^c}}$, defined in equations (3) and (4).

If we observe the results depicted in Table 1, we can draw some useful observations regarding the performance of Fuzzy ARTMAP under the three different modes of training.

1. The number of nodes created by Fuzzy ARTMAP trained under Mode 3 (cross-validated training) is significantly smaller than the number of nodes created by Fuzzy ARTMAP trained under Modes 1 (complete training) and 2 (one epoch of training). This observation is more pronounced for higher overlap datasets.

2. The generalization performance of Fuzzy ARTMAP trained under Mode 3 (cross-validated training) is better than the generalization performance of Fuzzy ARTMAP trained under Mode 1 (complete training) or Mode 2 (one epoch of training).

3. The difference in the generalization performance between Modes 3 (cross-validated training) and Mode 2 (one epoch of training) is larger than the difference in the generalization performance between Modes 3 and Mode 1 (complete training).

4. The difference in the number of nodes created between Modes 1 (complete training) and Mode 3 (cross-validated training) is larger than the difference in the number of nodes created between Modes 2 (one epoch of training) and Mode 1.

5. The above observations are valid for all the dimensions (2, 5, 10) and all the number of distinct classes (2, 3) that we experimented with.

4 Conclusions

In this paper we investigated the relative performance of Fuzzy ARTMAP trained to completion, or trained for 1 epoch, compared to the performance of Fuzzy ARTMAP trained until the maximum performance on a validation set is achieved. The results on the artificial databases, where we could control the amount of data used, the dimensionality of the input patterns and the degree of overlap of data belonging to different classes, indicate that cross-validation help us discover a Fuzzy ARTMAP network with increased generalization and significantly reduced number of nodes. These conclusions were more pronounced as we moved from databases of low overlap to databases of higher overlap. We have also conducted some experiments with the Nursery and Letters databases (extracted from the UCI repository; see Murphy et al., 1994) to investigate the issue of overtraining and the advantages of using cross-validation. The results are shown in Tables 2 and 3. Tables 2 and 3 indicate that for these two databases overtraining in Fuzzy ARTMAP is not an issue. But they also show that cross-validation is very useful in these databases because it identifies when the stopping of the training should occur.

5 References

[1] Carpenter, G. A., S. Grossberg, N. Markuzon, J. H. Reynolds and D. B. Rosen, "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multi-dimensional maps", *IEEE Transactions on Neural Networks*, Vol. 3, No. 5, 1992, pp. 698-713.

[2] Gomez Sanchez, E., Y. A. Dimitriadis, J. M. Cano Izquierdo, and J. Lopez Colorado, "MicroARTMAP: Use of mutual information for category reduction in Fuzzy ARTMAP", *Proceedings of the International Joint Conference in Neural Networks (IJCNN) 2000*, Como, Italy, July 24-28, 2000, pp. VI 47 -- VI 52.

[3] Kohavi, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection", *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 1137-1143.

[4] Murhpy, P., and D. Aha, UCI Repository of Machine Learning Databases, *Technical Report*, http://www.ics.edu/mlearn/MLRepository.html, Department of Computer Science, University of California Irvine, 1994.

[5] Vertzi, S., G. L. Heileman, M. Georgiopoulos, and M. J. Healy, "Boosting the performance of ARTMAP", *Proceedings of the 1998 International Joint Conference on Neural Networks (IJCNN-98)*, Anchorage, Alaska, June 1998, pp. 396-401.

[6] Williamson, J. R., "Gaussian ARTMAP: A neural network for fast incremental learning of noisy multi-dimensional maps", *Neural Networks*, Vol. 9, 1996, pp. 881-897.

Table 1: Comparison of Average Percentage of Correct Classification (*PCC*'s) and Average Node Compression Ratios (*CR*'s) for the three different Fuzzy ARTMAP training modes (1, 2, 3) and three degrees of overlap (low, medium, high) using artificial databases.

Overlap	*Classes*	*Dimension*	$\overline{PCC^{c}_{tes}}$	$\overline{PCC_{tes}} - \overline{PCC^{c}_{tes}}$	$\overline{PCC_{tes}} - \overline{PCC^{1EP}_{tes}}$	$\overline{CR^{c}}$	$\overline{CR^{1EP}}$
Low	2	2	95.39	1.11	1.82	44.21	14.94
	2	5	96.78	0.79	1.92	19.31	5.04
	2	10	99.76	0.08	0.43	3.26	2.02
	3	2	99.95	0.86	1.51	45.32	15.75
	3	*5*	*99.19*	*0.08*	*0.51*	*10.31*	*3.82*
	3	10	99.57	0.31	0.68	3.23	2.05
Medium	2	2	84.50	2.69	4.34	63.54	23.34
	2	5	83.03	0.29	2.44	42.98	10.87
	2	10	83.59	1.27	3.66	18.38	4.27
	3	2	85.22	2.31	4.20	75.19	28.55
	3	5	83.51	2.61	4.81	55.84	14.34
	3	***10***	***85.66***	***2.38***	***4.34***	***34.75***	***7.91***
High	2	2	70.34	2.53	3.96	44.97	18.22
	2	5	68.09	2.43	3.94	51.45	14.17
	2	10	68.05	2.73	4.24	28.97	6.89
	3	*2*	*67.22*	*3.02*	*4.95*	*91.00*	*43.71*
	3	5	63.61	2.24	3.90	93.10	28.90
	3	10	73.06	1.01	2.62	17.96	4.14

Table 2: Comparison of Percentage of Correct Classification (*PCC*'s) and Number of Nodes Created for the three different Fuzzy ARTMAP training modes (1, 2, 3) using the Nursery database.

Order	PCC_{v}	PCC_{tes}	PCC^{c}_{tes}	$PCC^{1\,EP}_{tes}$	PCC_{tr}	PCC^{c}_{tr}	$PCC^{1\,EP}_{tr}$	N_a	N^{c}_{a}	N^{1EP}_{a}
1	94.54	93.92	93.92	90.68	100	100	93.12	453	453	177
2	94.26	93.71	93.74	85.71	99.91	100	89.68	506	510	177
3	94.16	93.74	93.74	87.53	100	100	90.65	525	537	192
4	94.66	94.11	94.14	89.73	99.89	100	90.94	458	462	173
5	95.62	94.54	94.6	89.14	99.66	100	92.16	457	469	186
Avg	*94.65*	*94.00*	*94.03*	*88.56*	*99.89*	*100*	*91.31*	*479.8*	*486.2*	*181*

Table 3: Comparison of Percentage of Correct Classification (*PCC*'s) and Number of Nodes Created for the three different Fuzzy ARTMAP training modes (1, 2, 3) using the Letters database.

Order	PCC_{v}	PCC_{tes}	PCC^{c}_{tes}	$PCC^{1\,EP}_{tes}$	PCC_{tr}	PCC^{c}_{tr}	$PCC^{1\,EP}_{tr}$	N_a	N^{c}_{a}	N^{1EP}_{a}
1	82.99	82.83	82.83	81.01	100	100	96.05	710	710	659
2	83.13	82.77	82.77	80.26	100	100	95.3	678	678	619
3	83.31	83.27	83.49	81.49	99.35	100	96.11	667	684	632
4	83.51	83.41	83.43	80.99	99.99	100	95.56	691	695	638
5	84.11	83.03	83.03	80.32	100	100	95.5	712	712	651
Avg	*83.41*	*83.06*	*83.11*	*80.81*	*99.87*	*100*	*95.70*	*691.6*	*695.8*	*639.8*