# Pruning Fuzzy ARTMAP Using the Minimum Description Length Principle in Learning from Clinical Databases

Ten-Ho Lin and Von-Wun Soo
*Department of Computer Science*
*National Tsing-Hua University, Hsinchu, Taiwan, R.O.C.*
*E-mail: soo@cs.nthu.edu.tw*

## Abstract

*Fuzzy ARTMAP is one of the families of the neural network architectures based on ART(Adaptive Resonance Theory) in which supervised learning can be carried out. However, it usually tends to create more categories than are actually needed. This often causes the so called overfitting problem, namely the performance of the networks in test set is not monotonically increasing with the additional training epochs and category creation, for fuzzy ARTMAP. In order to avoid the overfitting problem, Carpenter and Tan [Carpenter and Tan, 1993] proposed a confidence-based pruning method by eliminating those categories that were either less useful or less accurate. This paper proposes yet another alternative pruning method that is based on the Minimal Description Length (MDL) principle. The MDL principle can be viewed as a tradeoff between theory complexity and data prediction accuracy given the theory. We adopted Cameron-Jones' error encoding scheme and Quinlan's modifier for theory encoding to estimate the fuzzy ARTMAP theory description length. A greedy search algorithm of the minimum description length to prune the fuzzy ARTMAP categories one by one is proposed. The experiments showed that fuzzy ARTMAP pruned with the MDL principle gave better performance with far fewer categories created than the original fuzzy ARTMAP and other machine learning systems on a number of benchmark clinical databases such as heart disease, breath cancer and diabetes databases.*
(**Subject Area:** Neural Networks; Knowledge Acquisition and Machine Learning)

## 1. Introduction

Learning and discovery from databases, or data mining, has recently raised much attention in both AI and database community [Frawley et al., 1991] [Agrawal and Psaila, 1995]. The main focus of the research is to induce regularities or rules using the databases as sources of training instances. The task is difficult in that the size of the databases can be potentially very huge and noise and missing data cannot be neglected. This is especially true for many clinical databases where patient records tended to be idiosyncrasy (exception) and imperfect (noisy or missing values).

Fuzzy ARTMAP [Carpenter et al., 1992] is a family of self-organized neural network architectures based on adaptive resonance theory (ART) [Carpenter and Grossberg, 1987] in which supervised learning can be carried out. Roughly speaking, the basic learning mechanism of fuzzy ARTMAP is that creating a new category (neuron unit) when an "unfamiliar" input instance is encountered while updating the connection weights of an old category when a "familiar" input instance with respect to the category is encountered. The level of familiarity is determined by thresholds of so called vigilance value and choice parameters in fuzzy ARTMAP. This learning mechanism makes fuzzy ARTMAP superior to other learning methods in that it can deal with both generalities and exceptions simultaneously.

However, fuzzy ARTMAP tended to create more categories than were actually needed. In many application domains, thousands of categories created in account for the input instances are not uncommon. This is an undesirable feature. Because it is difficult to interpret the learning results of fuzzy ARTMAP in terms of thousands of categories, not to mention that many categories created actually contribute nothing to the prediction accuracy. This belongs to the well

1082-3409/97 $10.00 © 1997 IEEE

known class of overfitting problems that more training efforts paid will not gain performance in testing. To avoid the overfitting problem, Carpenter and Tan [1993] introduced a pruning algorithm based on a so called confidence factor. The confidence factor of a category is a score in terms of a combination of its usage and performance accuracy. Their method not only reduced the network size to one third but also slightly improved the prediction accuracy in a diabetes database. However, the confidence-based pruning algorithm requires a separate training set (known as prediction set) to help learning, and although intuitively acceptable is basically a rule-of-thumb heuristic.

We consider creating a fuzzy ARTMAP category a tradeoff between theory complexity and performance accuracy, namely, creating a category can hopefully increase the performance accuracy but will also increase the complexity of the theory. The Minimum Description Length (MDL) principle [Rissanen, 1983] basically is a Occam's razor that can help to select among competing theories a balance between theory complexity and data prediction accuracy given the theory. The balance selected by the MDL principle is a bias toward a parsimony theory. Besides, the MDL principle has a profound root in information theory [Rissanen, 1983].

Hence we developed a pruning algorithm based on the MDL principle for fuzzy ARTMAP. To evaluate and compare the performance of the new learning scheme, we have done two experiments: 1) using a breast cancer database to compare the performance of the original fuzzy ARTMAP against that of fuzzy ARTMAP with MDL pruning. 2) using a Pima Indian diabetes database to compare the fuzzy ARTMAP performance of the confidence-based pruning against MDL pruning.

In the section 2, we will briefly describe the fuzzy ARTMAP architecture and its learning mechanism. In section 3, we discuss the minimum description length principle and how the it is applied to estimate the theory and data encoding in the fuzzy ARTMAP. In section 4, we describe the MDL-based pruning search algorithm for fuzzy ARTMAP. In section 5, we showed two experiments on different benchmark databases and compared their results. In section 6, we make discussion and conclusion.

## 2. The fuzzy ARTMAP architecture

Fuzzy ARTMAP is a neural network architecture that performs incremental supervised learning of recognition categories and multidimensional maps of both analog and binary patterns [Carpenter et al., 1992]. It consists of two fuzzy ART [Carpenter, Grossberg, and Rosen, 1991] modules, namely ART$_a$ and ART$_b$, linked via an inter-ART module, called a *map field* F$^{ab}$ as in Fig. 1. Each field in fuzzy ART, represented as a square in Fig. 1, consists of a set of neurons. The map field F$^{ab}$ links each category to its prediction. In classification tasks, ART$_b$ does nothing more than the identical mapping which directly maps the target vector b into the vector field F$_2^b$ and can be ignored. Each F$_2^a$ node corresponds to a category. The input vectors are preprocessed by a mechanism called *complement coding* where every input vectors is represented by a pair of (a, 1-a), namely, the input pattern a, and its complement 1-a. With the complement coding option, the weight vector w$_j^a$ of a category j can be viewed as a hyperrectangle over the input space under a geometric interpretation. Each category corresponds to a fuzzy inference rule, which tells a prediction is more possible when a input vector falls nearer (or within) the hyperrectangle.

The learning mechanism in fuzzy ARTMAP are conducted at ART$_a$ and map field F$^{ab}$. It first carries out a vigilance test as following: The category j with the highest value of choice function $T_j = | I \wedge w_j | / (\alpha + |w_j|)$ is chosen, where I is the input vector, $\wedge$ is the fuzzy AND operation, w$_j$ is the weight vector of category j and $\alpha$ is the choice parameter. The chosen category is then tested by the vigilance test $| I \wedge w_j | / | I | > \rho$, where $\rho$ is the vigilance parameter. If it fails the test, the category with the second high choice function is chosen and repeat the procedure above. If the chosen category which passes the vigilance test but predictes a wrong target class, the vigilance parameter is temporary increased and repeat the category choosing and vigilance test process. If the chosen category passes the vigilance test and predicts correctly, it will learn the input vector by updating the weight according to the following formula:

$$W_J^{(new)} = \beta(I \wedge W_J^{(old)}) + (1 - \beta) W_J^{(old)} \quad (1)$$

where $\beta$ is the learning rate. Finally, if no category passes all the tests above, a new category will be created to account for this particular input.

Proceedings of the 9th International Conference on Tools with Artificial Intelligence (ICTAI '97)
1082-3409/97 $10.00 © 1997 IEEE

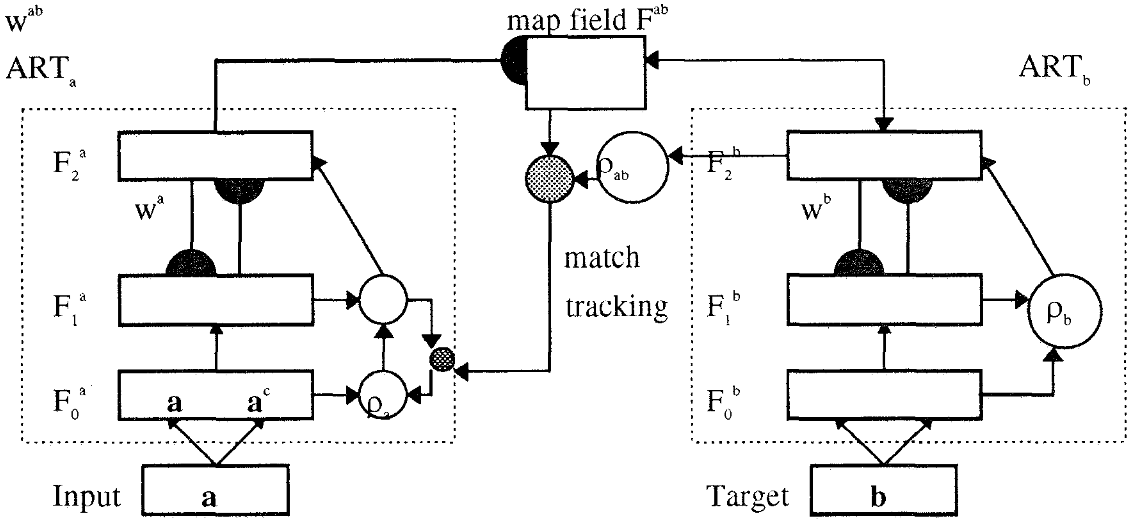

**Figure 1**: Fuzzy ARTMAP architecture [Carpenter et al., 1992]

## 3. The MDL principle and fuzzy ARTMAP pruning algorithm

When the learning is done for fuzzy ARTMAP, it enters the pruning phase where we need to decide which unnecessary category to prune away. We will discuss how minimum description length principle is used to guide pruning.

### 3.1. The Minimum Description Length

Let L(T) denote the number of bits needed to encode the theory T, and L(D|T) denote the number of bits needed to encode the data D with respect to T. L(T) measures the complexity of the theory, and L(D|T) measures how well the theory match the data (fewer bits indicate better fit). The Minimum Description Length principle states that the best theory is the one with the least number of bits required to encode the theory and the data given the theory. In other words, the best theory T is the one that minimizes

$$L(T) + L(D \mid T). \qquad (2)$$

The encoding length of a theory, L(T), for fuzzy ARTMAP is to be discussed in section 3.2 and 3.3. The encoding length of L(D|T) will be discussed in section 3.4.

### 3.2. Quantization levels of connection weights

To estimate the encoding complexity of theory, L(T), for fuzzy ARTMAP, one needs to know how a quantity of a weight is represented. Carpenter and Tan [1993] suggested the category weights be quantized in terms of nominal rather than numerical values. Quantization level Q is defined as the number of possible values for each category weight. For example, when Q=3, the possible weight values can be expressed as "low", "medium", or "high". There are two quantization methods: quantization by truncation and quantization by round-off. They were found to give similar performance. Quantization by round-off was used in this paper, which is described below:

If all possible values are $V_j = j / (Q-1)$, j=0, 1, ..., Q-1, reduce the weight w to the nearest $V_j$. Express it mathematically,

$$w^{(\text{quantized})} = \lfloor (Q-1)w + 0.5 \rfloor / (Q-1) \qquad (3)$$

### 3.3. The description length of Fuzzy ARTMAP

We restrict our research on classification problems, in which the binary target vectors consists of all 0's but only one 1. Hence the $ART_b$ module performs nothing but identical mapping and can be ignored. All we need to estimate is the description length of encoding the $ART_a$ module only. Since $ART_a$ consists of a collection of categories and each category is a vector of connection weight, the problem reduces to how to encode a connection weight. Assume the weights are quantized into Q values. With complement coding and the hyperrectangle interpretation, weights can be expressed in terms of a pair of the lower and upper vectors for categories. For a lower vector not equal to an upper vector, there are $\binom{Q}{2}$ possible combinations. For a lower vector equal to an upper vector, there are Q possible values. So, the description length of a weight is $\binom{Q}{2}+Q=\frac{Q(Q+1)}{2}$. A category consists of M pairs of weights, where M is the number of attributes.

$$M\log(\binom{Q}{2}+Q) \text{ bits.} \qquad (4)$$

Assume a fuzzy ARTMAP with N categories which is pruned from the original fuzzy ARTMAP with $N_{max}$ categories. First the number N should be encoded. Because N can be $1,2,\ldots,N_{max}$, it takes log $N_{max}$ bits to encode the number N. Second, to encode N categories, $N \times M \times \log(\binom{Q}{2}+Q)$ bits are needed. Finally, since the ordering of categories is unimportant, log N! is subtracted from the description length. Therefore, the overall theory coding length L(T) of a fuzzy ARTMAP network is

$$\log N_{max} + N \times M \times \log(\binom{Q}{2}+Q) - \log N!$$

(the length of encoding $N_{max}$ possible categories plus the total length of N categories subtract the information of encoding the ordering of N categories.)

Now we consider the unquantized weights which are difficult to encode because they are arbitrary real numbers between 0 and 1. However, we could assume they are quantized into W values, called *weight precision* which is a threshold parameter. If W is large, it is assumed that more information is on one weight so more categories should be pruned; if W is small, it is assumed that less information is on one weight so fewer categories will be pruned.

### 3.4. Quinlan's modifier

It was pointed out that traditional methods of applying MDL could lead to poor accuracy in categorical domain [Quinlan, 1994]. Quinlan proposed a modifier to remedy the problem. Quinlan's modifier is a mechanism that improve the MDL selection by restricting candidate theories that tends to perform poorly. Let the probability, or proportion of positives be P(D+), the probability (proportion) of data instances predicted as positives be P(D+|T). The idea is that P(D+|T) should fall between P(D+) ± s.d.(P(D+)), where s.d.(P(D+)) is the standard deviation of P(D+). Assume data instances are independent, and the probability of each data instance being positive is identical. Then the number of positives conforms to the binomial model, so

$$s.d.(P(D+)) = \frac{\sqrt{P(D+)(1-P(D+))}}{|D|}$$

(5)

The restricting scheme is as follow: let

$$V = |P(D+|T) - P(D+)| \,/\, s.d.(P(D+)). \qquad (6)$$

If V>1, P(D+|T) is not between P(D+) ± s.d.(P(D+)) and the theory is not likely to match the data, the theory is modified by $\sqrt{V}$ as a penalty. The expression for L(T) above is multiplied by Quinlan's modifier. In one equation, the theory description length of a fuzzy ARTMAP is

$$L(T) = \max(\sqrt{\frac{|P(D+|T)-P(D+)|}{s.d.(P(D+))}},1)$$
$$\times (\log(N_{max}) + NM\log(\binom{Q}{2}+Q) - \log N!)$$

(7)

### 3.4. The description length of data given a theory

In two-class classification problems, encoding data given a theory, namely L(D|T), is equivalent to encoding the errors of the theory [Quinlan, 1994]. We could, therefore, encode the L(D|T) in terms of a function of prediction errors of a theory. However, as pointed out by Quinlan [1994] most description length functions are not monotonically increasing with respect to number of errors. This is because if more than half of data was falsely predicted in two class problems, predictions can be made on the wrong side of the learning system.

1082-3409/97 $10.00 © 1997 IEEE

**Table 1**: The experimental results on Wisconsin breast cancer database

| Learning systems | Train/Test size | Accuracy | Rules or Instances |
|---|---|---|---|
| C4.5 + MDL | 629/70 | 95.5% | 8.5 |
| Fuzzy ARTMAP | 629/70 | 91.86% | 13.0 |
| Fuzzy ARTMAP + MDL | 629/70 | 95.57% | 4.2 |

This is an undesirable property because it may prefer a less accurate theory to a more accurate theory when their complexities are the same. Cameron-Jones [1992] has proposed an error encoding scheme that increases monotonically with e, the number of errors. It first defines a code for e=0, then for e=1, 2, etc. Each of these codes is defined by an integer starting with 1 but, since we want no differentiation between theories with the same number e of errors, the encoding length for each theory in the group of the same number of errors is taken to be that of the integer associated with the last theory in the group.

$$\sum_{i=0}^{e}\binom{|D|}{i} \qquad (7)$$

where |D| is the total number of data instances.

To encode this integer, an encoding scheme for arbitrary integer is needed. Rissanen [1983] has claimed that it need

$\log^*(n) + C$ bits

to encode an arbitrary integer n, where C is a constant (and does not matter what number we use in pruning) and

$$\log^*(n) = \log(n)+\log(\log(n))+\log(\log(\log(n)))+\ldots$$

The recursive logarithm function for log*(n) is repeated until the term becomes negative.

Hence the Cameron-Jones' error encoding cost for e errors is

$$\log^*\left(\sum_{i=0}^{e}\binom{|D|}{i}\right)+C \text{ bits.} \qquad (8)$$

## 4. Searching for the theory with the minimum description length

The number of possible pruned networks from a fuzzy ARTMAP with N categories is equal to the number of nonempty subsets of N categories, $2^N$ - 1. This makes exhaustive search of the network with minimum description length almost impossible (there may be thousands of categories in a complex fuzzy ARTMAP, for example, letter recognition benchmark problem in [Carpenter et al., 1992]). We used a greedy algorithm, or a best-first search to search the theory with minimum length. Although this search strategy may found local minimum instead of global minimum, it is a feasible choice due to the time constraint. The search algorithm is described below:

1.Find the description length of the fuzzy ARTMAP $L_0$.
2.For each category j, assume it is pruned, and find the description length L'(j) according to expression (7) and (9).
3.Find J such that L'(J) = min(L'(j)), i.e. find the category J that decreases the description length most by its pruning.
4.If L'(j) ≥ $L_0$, i.e. pruning can no longer decrease description length, or the number of categories are 2, algorithm ends.
5.If L'(j) < $L_0$, prune category J, i.e. prune the corresponding $F_2^a$ node J.
6.Repeat 1 until algorithm ends in step 4.

## 5. Experimental results

### 5.1. Wisconsin breast cancer database

The Wisconsin breast cancer database, available via UCI machine learning repository [Murphy and Aha, 1992], was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. The task is to classify between benign and malignant cases. This database contains 699 instances, of which 458 (66.5%) are benign and 241 (34.5%) are malignant, and 16 missing values. We chose 9 out of 10 attributes (ID number removed) for learning. A fuzzy ARTMAP with missing value treatment using the complement coding strategy can be found in the source program for [Carpenter, 1993]: An input vector **a** = $(a_1,\ldots,a_M)$ is turned to **I** = (**a**, **a**$^c$) = $(a1,\ldots,a_M,a_1^c,\ldots,a_M^c)$ and sent to fuzzy ARTMAP. For known value $a_i$, $a_i^c = 1-a_i$. For missing value, let $a_i = a_i^c = 1$. This is equivalent to assuming the unknown value fit in the range of every category in that attribute.

Quinlan [1995] used C4.5 with (biased) MDL

Proceedings of the 9th International Conference on Tools with Artificial Intelligence (ICTAI '97)
1082-3409/97 $10.00 © 1997 IEEE

**Table 2**: PID database experiments results (Note that $\alpha$ is the choice parameter; $\rho$ is the vigilance threshold; Q is the quantization level; the question mark ? denotes data are not available; — denotes the parameter is irrelevant or note used; and W is the weight precision. The ##(*20) denotes that the data ## is averaged over 20 voters)

| Learning system | $\alpha$ | $\rho$ | Q | Accuracy | # of rules or nodes |
|---|---|---|---|---|---|
| ADAP | — | — | — | 76.0 | 100000 |
| C4.5 + MDL | — | — | — | 72.4 | 13.3 |
| Fuzzy ARTMAP | ? | ? | — | 75.9 | 63.5 (*20) |
| confidence pruning | ? | ? | — | 78.5 | 19.6 (*20) |
| confidence pruning | ? | ? | 5 | 77.5 | 19.6 (*20) |
| Fuzzy ARTMAP | 0.2 | 0.0 | — | 73.6 | 75.29 (*20) |
| confidence pruning | 0.2 | 0.0 | — | 75.4 | 18.86 (*20) |
| MDL pruning, W=5 | 0.2 | 0.0 | — | 75.11 | 9.90 (*20) |
| confidence pruning | 0.2 | 0.0 | 5 | 72.1 | 19.90 (*20) |
| MDL pruning | 0.2 | 0.0 | 5 | 72.70 | 8.20 (*20) |
| confidence pruning | 0.2 | 0.0 | 6 | 72.2 | 19.08 (*20) |
| MDL pruning | 0.2 | 0.0 | 6 | 74.79 | 8.27 (*20) |
| Fuzzy ARTMAP | 0.05 | 0.6 | — | 73.0 | 53.91 (*20) |
| confidence pruning | 0.05 | 0.6 | — | 72.0 | 14.56 (*20) |
| MDL pruning, W=5 | 0.05 | 0.6 | — | 74.74 | 9.68 (*20) |
| confidence pruning | 0.05 | 0.6 | 5 | 71.5 | 14.91 (*20) |
| MDL pruning | 0.05 | 0.6 | 5 | 70.16 | 7.33 (*20) |
| confidence pruning | 0.05 | 0.6 | 6 | 72.5 | 14.65 (*20) |
| MDL pruning | 0.05 | 0.6 | 6 | 71.36 | 7.74 (*20) |

pruning and achieved 95.5% accuracy by 8.5 rules with 629 training data and 70 testing data. Quinlan used the updated version of data that we used. Another fuzzy ARTMAP simulation with $\alpha$=0.01, $\beta$=1, $\rho$=0, with the same number of training and test data, was run for 10 trials. It gave 91.86% accuracy by 13.0 rules on average, which is worse than C4.5 + MDL in both performance and number of rules. After MDL pruning with weight precision 5, the accuracy increased to 95.57% by 4.2 rules, which is similar to C4.5 + MDL pruning in accuracy but used about half as many rules. The experimental results are summarized in Table 1.

### 5.2. Pima Indian Diabetes database

The Pima Indian Diabetes (PID) database, also obtained from UCI machine learning repository is originally owned by National Institute of Diabetes and Digestive and Kidney Diseases. It contains 768 instances, 8 input attributes and 1 target, which represents whether the data shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl found at any survey examination or during routine medical care). 268 instances of the data are positive, which is 34.9% of the database. There is no missing value instance.

ADAP, a feedforward neural network model, has been applied to PID database using 576 training data and 192 testing data and achieved 76% accuracy by 100000 association units [Smith et al., 1988]. All the following tests used the same training size. The prediction based on 20 voting fuzzy ARTMAP's achieved 75.9% accuracy with average 63.5 (*20) rules, which is very low comparing to 100000 association units of ADAP. With confidence-based pruning, the accuracy could be improved to 78.5% using even fewer rules, 19.6 (*20) [Carpenter, 1993]. Unfortunately, the exact parameters were not listed in the paper, and we couldn't repeated the above results.

In order to compare MDL-based pruning with confidence-based pruning, both techniques were tested under two sets of parameters, $\alpha$=0.2, $\beta$=1, $\rho$=0 and $\alpha$=0.05, $\beta$=1, $\rho$=0.6, and different quantization levels. The two sets of parameters have different vigilance values and different number of categories. All simulations are repeated 10 times, with 20 voters. Under these parameters and quantization levels, MDL-based pruning gave slightly better accuracy in about half as many rules. Quantization level 6 gave better accuracy and fewer rules than quantization level 5 in both

1082-3409/97 $10.00 © 1997 IEEE

pruning methods. The results can be found in Table 2.

# 6. Conclusion

## 6.1. Summary and discussion

Fuzzy ARTMAP is a powerful neural network model with many useful characteristics, including stability, guaranteed convergence, and online learning. Overfitting avoidance techniques in fuzzy ARTMAP are very important but were seldom addressed.

Fuzzy ARTMAP gave better performance and fewer rules over other machine learning algorithms and neural network models in different benchmark databases learning problems. We found that Fuzzy ARTMAP with MDL-based pruning used fewer categories and often even better performance than fuzzy ARTMAP without MDL pruning. We also found that MDL-based pruning extracted fewer rules and often better accuracy than confidence-based pruning, and need not a separate predicting set.

However, our current greedy search strategy in search of shortest description length theory seems to be problematic in that it may trap in a local optimal. This is due to the execution time of MDL-based pruning. Because computing description length needs to go through all training instances, it takes about the time of a whole training epoch to pruning a single category. In PID database, the original fuzzy ARTMAP has more than 60 categories and the pruned network has less than 10 categories, which means more than 50 categories are pruned, and about the time of 50 epochs is required. It takes a large amount of time in comparison to 10 to 20 epochs that fuzzy ARTMAP usually takes for training. The confidence-based pruning only needs to go through the training data once. This is the reason it didn't look over more thorough search space of all possible networks. However, since fuzzy ARTMAP pruning requires off-line learning, the execution time is not so critical.

## 6.2. Future work

1. The MDL encoding schemes and Quinlan's modifier are developed based on two-classes (positive and negative) classification. We wish to generalize them to more-than-two classes in the future.
2. The nominal and missing value treatments in fuzzy ARTMAP also need further study.
3. One important and useful property of fuzzy ARTMAP is online learning. Both fuzzy ARTMAP pruning techniques conflict with the concept of online learning, because pruning must be applied after the network is trained. Is it possible for a MDL-based online checking criteria to guide the online learning system in category creation and overfitting avoidance? This is definitely an interesting open problem.

### Acknowledgment

This research is supported by the collaboration research project of National Tsing Hua University and Veteran General Hospital under the Grant No. VGHTH 85-008-2.

### Reference

Agrawal, R. and Psaila, G. (1995). Active data mining. *Proceedings, First International Conference on Knowledge Discovery and Data Mining*, 216-221. AAAI Press.

Cameron-Jones, R. M. (1992). Minimum description length instance-based learning. *Proceedings 5th Australian Joint Conference on Artificial Intelligence*, (A. Adams and L. Sterling, Eds), Singapore: World Scientific, 368-373.

Carpenter, G. A., and Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37, 54-155.

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

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

Carpenter, G. A. and Tan, A. H. (1993). Rule extraction, Fuzzy ARTMAP, and medical databases. *Proceedings, World Congress on Neural Networks, Portland, OR* (Vol. I, pp.501-506). Hillsdale, NJ: Lawrence Erlbaum Associates.

Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1991). Knowledge discovery in databases: an overview. *Knowledge Discovery in Databases*, 71-92. AAAI Press / The MIT Press

Gallager, R. G. (1968). *Information Theory and Reliable Communication*. New York: Wiley.

Gennari, J. H., Langley, P., and Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence*, 40, 11-61.

Murphy, P. M. and Aha, D. W. (1992) *UCI Repository of machine learning databases* [Machine readable data repository]. Irvine, CA: University of California, Department of Information and Computer Science.

Quinlan, J. R. (1994). The Minimum Description Length principle and categorical theories. *Proceedings 11th International Conference on Machine Learning*, New Brunswick, 233-241. San Francisco: Morgan Kaufmann.

Quinlan, J. R. (1995). MDL and categorical theories (continued). *Proceedings 12th International Conference on Machine*, 464-469. San Francisco: Morgan Kaufmann.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11, 416-431.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., and Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Symposium on Computer Applications and Medical Care*, 261-265. IEEE Computer Society Press.

Zhang, J. (1992). Selecting typical instances in instance-based learning. *Proceedings of the Ninth International Machine Learning Conference*, 470-479. Aberdeen, Scotland: Morgan Kaufmann.